

Joint Modeling of Big Five and HEXACO for Multimodal Apparent Personality-trait Recognition

Ryo Masumura, Shota Orihashi, Mana Ihori, Tomohiro Tanaka, Naoki Makishima, Taiga Yamane,
Naotaka Kawata, Satoshi Suzuki, Taichi Katayama
NTT, Inc., Japan
E-mail: ryo.masumura@ntt.com

Abstract—This paper proposes a joint modeling method of the Big Five, which has long been studied, and HEXACO, which has recently attracted attention in psychology, for automatically recognizing apparent personality traits from multimodal human behavior. Most previous studies have used the Big Five for multimodal apparent personality-trait recognition. However, no study has focused on apparent HEXACO which can evaluate an *Honesty-Humility* trait related to displaced aggression and vengefulness, social-dominance orientation, etc. In addition, the relationships between the Big Five and HEXACO when modeled by machine learning have not been clarified. We expect awareness of multimodal human behavior to improve by considering these relationships. The key advance of our proposed method is to optimize jointly recognizing the Big Five and HEXACO. Experiments using a self-introduction video dataset demonstrate that the proposed method can effectively recognize the Big Five and HEXACO.

I. INTRODUCTION

Recognizing people's personality traits has been a central topic in the psychological and engineering fields. Two types of personality traits have been considered; self-assessed and apparent perceived by observers. In psychology, personality traits are measured through questionnaire-based personality tests for both the self-assessed and apparent personality traits. While the personality test results for self-assessed personality traits can be attained from one self-trial, those for apparent personality traits need to be judged by many other people. To recognize the apparent personality traits without the help of people other than oneself, researchers in the engineering field have studied multimodal apparent personality-trait recognition in which apparent personality traits are automatically recognized from multimodal human behavior using machine learning [1]–[4].

Many modeling methods for multimodal personality-trait recognition have been studied. Deep-learning-based methods for learning effective representations from multimodal human behavior without introducing hand-crafted features are now widely used [5]–[9]. With these methods, personality traits are estimated by integrating speech, visual, and text information exploited from human behavior. When modeling apparent personality traits, most studies modeled to recognize the Big Five personality traits of *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism* [10], [11]. However, no study has focused on recognizing apparent personality traits other than the Big Five. This is because most datasets were developed for measuring the Big Five [2].

In this study, we focus on the HEXACO traits [12], [13] supported by recent theoretical and empirical studies on alternatives to the Big Five. HEXACO is a six-factor framework that includes *Honesty-Humility* and variants of the Big Five traits, i.e., *Emotionality*, *Extraversion*, *Agreeableness*, *Conscientiousness*, and *Openness*. It has been investigated that *Honesty-Humility* is strongly negatively correlated with a variety of factors (e.g., displaced aggression and vengefulness [14], social-dominance orientation [15], and workplace misconduct [16]) and has little correlation with the Big Five traits, so it would be worthwhile to automatically recognize the apparent HEXACO personality traits from multimodal human behavior. Note that there was one trial that examined self-reported HEXACO traits from social-media text posts [17], but inferring apparent observer-perceiving HEXACO traits from multimodal human behavior has not been investigated. In addition, the relationships between the Big Five and HEXACO when modeled by machine learning have not been clarified, although their relationships have been analyzed from many psychological aspects. For example, characteristics other than *Honesty-Humility* in HEXACO are closely related to the corresponding characteristics in the Big Five [18], [19]. It has also been shown that *Honesty-Humility* is partially related to *Agreeableness* of the Big Five [20]. By modeling multimodal personality-trait recognition that can take into account these relationships, we expect to promote robustness to being aware of various multimodal human behaviors.

To explicitly consider the relationships between the Big Five and HEXACO, we propose a joint-modeling method of the Big Five and HEXACO for multimodal apparent personality-trait recognition. Our proposed method simultaneously optimizes recognizing the Big Five and HEXACO from multimodal audio-video information. We model them using a multimodal-transformer architecture [21] to increase the awareness of multimodal human behavior in the Big Five and HEXACO. For this modeling, we extend an existing self-introduction video dataset [22] by assigning not only the Big Five and HEXACO. Our dataset consists of 50 Big Five questionnaire items [23], [24] and 60 HEXACO questionnaire items [25] collected from five observers of over 10,000 self-introduction videos. In experiments using the dataset, we show that joint modeling can improve the recognition performance of Big Five and HEXACO compared with individual modeling.

Our contributions are summarized as follows.

TABLE I
A 60-ITEM HEXACO QUESTIONNAIRE.

id	key	question
1.	O-	He/she would be quite bored by a visit to an art gallery.
2.	C+	He/she plans ahead and organizes things, to avoid scrambling at the last minute.
3.	A+	He/she rarely holds a grudge, even against people who have badly wronged him/her.
4.	X+	He/she feels reasonably satisfied with himself/herself overall.
5.	E+	He/she would feel afraid if he/she had to travel in bad weather conditions.
6.	H+	He/she wouldn't use flattery to get a raise or promotion at work, even if he/she thought it would succeed.
...
55.	O-	He/she finds it boring to discuss philosophy.
56.	C-	He/she prefers to do whatever comes to mind, rather than stick to a plan.
57.	A-	When people tell him/her that he/she is wrong, his/her first reaction is to argue with them.
58.	X+	When he/she is in a group of people, he/she is often the one who speaks on behalf of the group.
59.	E-	He/she remains unemotional even in situations where most people get very sentimental.
60.	H-	He/she'd be tempted to use counterfeit money, if he/she were sure he/she could get away with it.

- This paper is the first to examine multimodal apparent personality-trait recognition involving HEXACO.
- This paper provides a joint modeling method of the Big Five and HEXACO, which yields the improved recognition performance of both traits.
- This paper is the first to investigate the relationships between the Big Five and other personality traits, i.e., HEXACO, in multimodal apparent personality-trait recognition.
- This paper presents a self-introduction video dataset to which the Big Five and HEXACO traits are jointly assigned by others.

II. DATASET

This section details our self-introduction video dataset.

A. Self-introduction Videos

We extended an existing self-introduction video dataset [22] by assigning not only the Big Five and HEXACO. The dataset includes 10,100 self-introduction videos collected from 1,010 participants. The following interview items are on the theme of self-introduction. “Please tell us about your hobbies.” “Please tell us about your favorite food.” “Please tell us about your favorite celebrity.” “Please tell us about the tourist spots that you are glad you visited.” “Please tell us about your most impressive childhood memories.” “Please tell us about some interesting people you have met.” “Please tell us about your favorite season.” “Please tell us about the place you would like to visit.” “Please tell us about something you would like to try.” “Please tell us about something you are not good at”. Ten videos were recorded from each participant, who were all Japanese. The recorded videos are composed of about 12,395 min of recordings, and the average duration of each video is 73.6 s. The maximum and minimum duration of all videos are 102.1 and 59.1 s, respectively. All videos were recorded using Zoom on laptop PCs. We recorded the videos at 25 fps in 1280 × 720 resolution. Camera views were frontal, and we recorded the upper part of the body. The audio was recorded at 16 kHz. We split the dataset into a training dataset containing 9,030 videos recorded from 903 participants, validation dataset

containing 500 videos recorded from 50 participants, and test dataset containing 570 videos recorded from the remaining 57 participants.

B. Annotations of Big Five and HEXACO

All recorded videos were annotated with apparent personality traits. We used the Big Five [10], [11] (*Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism*) and HEXACO [12], [13] (*Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness*) for the apparent personality traits. To annotate people's apparent personality traits, we recruited 200 observers who did not know the 1,010 participants. We used a 50-item Big Five questionnaire [23], [24] and 60-item HEXACO questionnaire [25]. The videos in the training and validation datasets were scored by five randomly selected observers and those in the test dataset were scored by ten randomly selected observers. In the test dataset, five annotations assigned ground-truth information, and the other five conducted human evaluation. Each observer watched each recorded video two or three times and answered the questionnaire. We used a five-point scale for scoring. Table 1 shows the 12 items in the 60-item HEXACO questionnaire. Each key in Table 1 represents which personality traits it pertains to. “H”, “E”, “X”, “A”, “C” and “O” represent *Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness*, respectively. For “+” keyed items, the response “Very Inaccurate” is assigned a value of 1, “Moderately Inaccurate” a value of 2, “Neither Inaccurate nor Accurate” a 3, “Moderately Accurate” a 4, and “Very Accurate” a value of 5. For “-” keyed items, the response “Very Inaccurate” is assigned a value of 5, “Moderately Inaccurate” a value of 4, “Neither Inaccurate nor Accurate” a 3, “Moderately Accurate” a 2, and “Very Accurate” a value of 1. Note that the annotators were instructed to avoid assigning “Neither Inaccurate nor Accurate” as much as possible. Once scores are assigned for all of the items in the scale, all the values are averaged to obtain a total scale score. Figures 1 and 2 respectively show the histograms of the annotated Big Five and HEXACO personality traits of our recorded videos. The scores of individual personality traits are

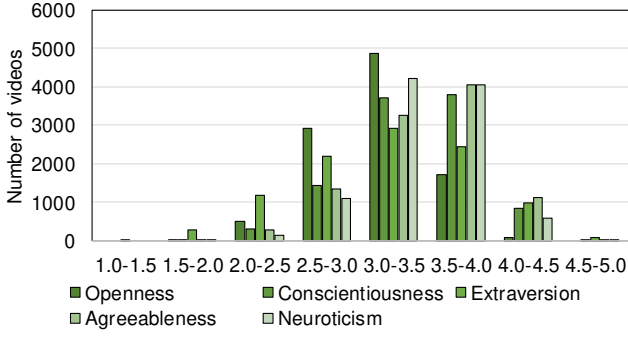


Fig. 1. The histograms of the annotated Big Five

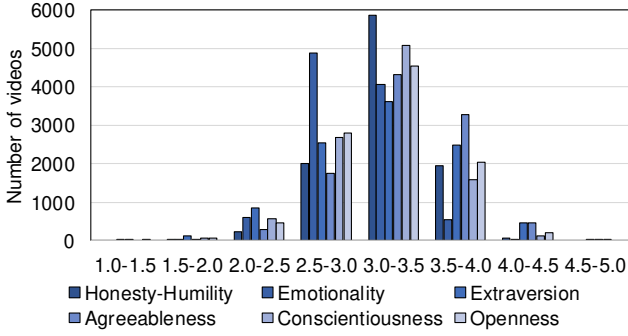


Fig. 2. The histograms of the annotated HEXACO

in the range of [1, 5]. Note that these scores are normalized in the range of [0, 1] when using deep-learning-based modeling methods.

III. JOINT MODELING OF BIG FIVE AND HEXACO WITH MULTIMODAL TRANSFORMER

This section details a joint modeling method of the Big Five and HEXACO.

A. Definitions

In this task, the Big Five scores $\hat{\mathbf{y}} = [\hat{y}_1, \dots, \hat{y}_5]^\top$ and HEXACO scores $\hat{\mathbf{z}} = [\hat{z}_1, \dots, \hat{z}_6]^\top$ are jointly estimated from an audio-visual video input, which is represented as audio features \mathbf{S} and their corresponding visual features \mathbf{U} . Audio features are generally extracted from speech information, and visual features are extracted from human RGB images. When modeling multimodal fine-grained apparent-personality-trait recognition, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ are estimated using

$$\{\hat{\mathbf{y}}, \hat{\mathbf{z}}\} = \mathcal{F}(\mathbf{S}, \mathbf{U}; \Theta), \quad (1)$$

where $\mathcal{F}(\cdot)$ is the model function and Θ represents the trainable model-parameter set. In addition, an automatic speech recognition (ASR) system can be used to convert the \mathbf{S} into text \mathbf{W} .

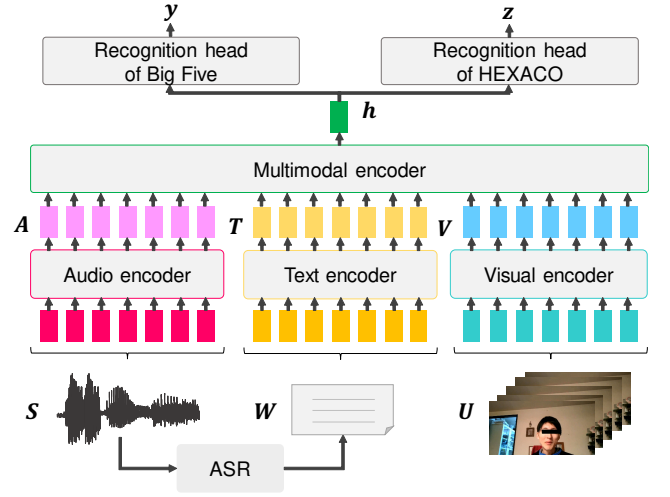


Fig. 3. Joint modeling of Big Five and HEXACO with multimodal transformer

B. Joint Modeling

Our proposed method uses a multimodal transformer architecture to effectively capture multimodal information. The advantage of this is that different types of features can be handled with the same input method. The architecture consists of four encoders: audio, text, visual, and multimodal. Figure 3 shows the architecture. The audio encoder converts audio features \mathbf{S} into audio representations \mathbf{A} , the text encoder converts text \mathbf{W} into text representations \mathbf{T} , and the visual encoder converts visual features \mathbf{U} into visual representations \mathbf{V} .

The multimodal encoder handles cross-modal interactions of outputs from the audio, text, and visual encoders. The inputs for the multimodal encoder are

$$\mathbf{H}_0 = \begin{cases} \text{TemporalConcat}(\mathbf{A}, \mathbf{T}, \mathbf{V}) & \text{if ASR is performed,} \\ \text{TemporalConcat}(\mathbf{A}, \mathbf{V}) & \text{else,} \end{cases} \quad (2)$$

$$\mathbf{H}'_0 = \text{AddSegment}(\mathbf{H}_0; \theta_{\text{segment}}), \quad (3)$$

where $\text{TemporalConcat}()$ is a function that concatenates inputs on the temporal axis, $\text{AddSegment}()$ is a function that adds a continuous vector in which modal-specific segment information is embedded to distinguish the concatenated vectors, and $\theta_{\text{segment}} \in \Theta$ are the trainable parameters. We obtain hidden vectors \mathbf{H} by

$$\mathbf{H} = \text{TransformerEnc}(\mathbf{H}'_0; \theta_{\text{multi}}), \quad (4)$$

where $\text{TransformerEnc}()$ is a function of the transformer encoder blocks [26] and $\theta_{\text{multi}} \in \Theta$ are the trainable parameters of the multimodal encoder. Note that the length of \mathbf{H} changes depending on the inputs.

Attentive pooling converts variable length \mathbf{H} into a fixed size vector. The fixed vector is obtained by

$$\mathbf{h} = \text{AttentivePooling}(\mathbf{H}; \theta_{\text{pool}}), \quad (5)$$

where $\theta_{\text{pool}} \in \Theta$ are the trainable parameters of attentive pooling, and $\text{AttentivePooling}()$ is the attentive-pooling function.

This model jointly estimates the Big Five and HEXACO scores by providing two prediction heads calculated as

$$\hat{z} = \text{Sigmoid}(\mathbf{h}; \theta_{\text{head}}^z), \quad (6)$$

$$\hat{y} = \text{Sigmoid}(\mathbf{h}; \theta_{\text{head}}^y), \quad (7)$$

where $\{\theta_{\text{head}}^z, \theta_{\text{head}}^y\} \in \Theta$ are the trainable parameters.

C. Training

To train Θ , we use a dataset of audio-visual video input, which is expressed as

$$\mathcal{D} = \{(\mathbf{S}^1, \mathbf{U}^1, \mathbf{y}^1, \mathbf{z}^1), \dots, (\mathbf{S}^{|\mathcal{D}|}, \mathbf{U}^{|\mathcal{D}|}, \mathbf{y}^{|\mathcal{D}|}, \mathbf{z}^{|\mathcal{D}|})\}. \quad (8)$$

Our joint model is trained with the mean absolute error loss between the ground-truth Big Five and estimated Big Five, and the mean absolute error loss between the ground-truth HEXACO and estimated HEXACO as

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} |\hat{\mathbf{y}}^d - \mathbf{y}^d| + \frac{1}{|\mathcal{D}|} \sum_{d=1}^{|\mathcal{D}|} |\hat{\mathbf{z}}^d - \mathbf{z}^d|. \quad (9)$$

By taking into account the relationships between the Big Five and HEXACO, we expect to promote robustness to being aware of various multimodal human behaviors.

IV. EXPERIMENTS

We used our dataset in the following experiments. We verified the effectiveness of our proposed joint-modeling method. We also investigated the relationships between the Big Five and HEXACO in multimodal apparent personality-trait recognition.

A. Setups

In our evaluation, we constructed two task-specific models, i.e., Big Five model and HEXACO model, and a joint model using a multimodal transformer architecture.

We carried out pre-processing to extract audio and visual features from video input. We extracted 80 log Mel-scale filterbank coefficients for the acoustic features, and the frame shift was 10 ms. Face regions in each input frame were detected with CenterNet [27] trained on the Wider Face dataset [28] for the visual features. The face images were cropped and resized to 128×128 , and down-sampled to 3 fps. We converted the audio features into text using a transformer-based end-to-end automatic-speech-recognition (ASR) system trained with 20K hrs of Japanese speech. The configuration was as follows. For the audio encoder, audio features passed two convolution and max-pooling layers with a stride of 2, so we down-sampled them to 1/4 along with the time axis. We stacked four transformer-encoder blocks. For the visual encoder, the convolutional-neural-network function was composed of the MobileNetV3 architecture [29], and two transformer encoder blocks were additionally stacked. We stacked six transformer-encoder blocks for the text encoder and two transformer-encoder blocks for the multimodal encoder. For each encoder, the dimensions of the output continuous representations were

set to 256, dimensions of the inner outputs were set to 1024, and number of heads in the multi-head attentions was set to 4. Swish activation was used for these encoders. For each prediction head, a fully connected layer with the sigmoid-activation function was used.

We pre-trained the parts of the multimodal transformer architecture. The audio encoder was pre-trained with masked prediction of hidden units [30] using over 20K hrs of Japanese speech. The text encoder was pre-trained with a masked language-modeling task [31] using over 100G tokens of text. The visual encoder was pre-trained with a still-image-based facial-expression-recognition task using RAF-DB [32] and AffectNet [33] datasets. Note that these pre-trained parameters were not frozen in the following main training. After the pre-training, all parameters in each model were trained. The mini-batch size was set to 8, and the dropout rate in the transformer blocks was set to 0.1. We used RAdam [34] for optimization. The training steps were stopped on the basis of early stopping using the validation dataset. We trained all models with one NVIDIA A6000 GPU.

B. Evaluation metrics

We evaluated task-specific models and a joint model in terms of Pearson's correlation coefficient and accuracy. The accuracy was computed in the same manner as with ChaLearn first impression [35], [36]. The accuracy for the k -th personality trait against the D test samples is defined as

$$\text{Accuracy}_k = 1 - \frac{1}{D} \sum_{d=1}^D |\hat{y}_k^d - y_k^d|, \quad (10)$$

where \hat{y}_k^d and y_k^d are the ground-truth and predicted scores of the k -th personality trait for the d -th test sample. Note that the scores were normalized in the range of [0, 1].

C. Results

Tables 1 and 2 show the multimodal apparent-personality-trait recognition performance for the Big Five and HEXACO, respectively. The experimental results show that audio features were more effective than visual features, and the visual features are comparatively effective in recognizing *Agreeableness* in the Big Five, *Emotionality* and *Agreeableness* in the HEXACO. In addition, combining audio, text, and visual inputs was effective for both the Big Five and HEXACO. This indicates that a multimodal transformer architecture with pre-trained encoders was effective in integrating multimodal information. The experimental results also show that the joint model outperformed the task-specific models for Big Five and HEXACO in most cases. This suggests that we can promote robustness to being aware of various multimodal human behaviors by explicitly taking into account the relationships between the Big Five and HEXACO. The highest performance was achieved by the joint model with audio, visual, and text inputs for the Big Five and HEXACO evaluation. The automatic recognition performance competed with human evaluation performance.

TABLE II
EXPERIMENTAL RESULTS OF RECOGNIZING BIG FIVE IN TERMS OF PEARSON'S CORRELATION COEFFICIENT (CORR.) AND ACCURACY (ACC.)

Modeling method	Input modals	<i>Openness</i>		<i>Conscientiousness</i>		<i>Extraversion</i>		<i>Agreeableness</i>		<i>Neuroticism</i>	
		Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.
Big Five model	Audio	0.493	93.9	0.604	93.2	0.647	91.2	0.572	92.3	0.473	93.5
Joint model	Audio	0.542	94.4	0.614	93.3	0.707	91.6	0.576	93.4	0.530	93.8
Big Five model	Visual	0.233	93.1	0.310	90.8	0.264	86.4	0.433	92.4	0.233	93.1
Joint model	Visual	0.228	92.9	0.332	91.2	0.315	87.2	0.452	92.6	0.286	93.3
Big Five model	Audio, Visual	0.544	94.4	0.604	93.5	0.735	91.0	0.615	92.6	0.532	94.0
Joint model	Audio, Visual	0.557	94.5	0.617	93.3	0.743	92.0	0.628	93.8	0.538	94.2
Big Five model	Audio, Visual, Text	0.585	94.6	0.675	93.8	0.752	92.4	0.617	92.7	0.586	94.1
Joint model	Audio, Visual, Text	0.595	94.8	0.686	93.9	0.757	92.6	0.657	94.0	0.586	94.2
<i>Human evaluation</i>		0.544	92.9	0.668	92.7	0.770	91.7	0.645	92.4	0.532	92.1

TABLE III
EXPERIMENTAL RESULTS OF RECOGNIZING HEXACO IN TERMS OF PEARSON'S CORRELATION COEFFICIENT (CORR.) AND ACCURACY (ACC.)

Modeling method	Input modals	<i>Honesty-Humility</i>		<i>Emotionality</i>		<i>Extraversion</i>		<i>Agreeableness</i>		<i>Conscientiousness</i>		<i>Openness</i>	
		Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.
HEXACO model	Audio	0.468	95.1	0.626	95.3	0.616	92.7	0.468	94.0	0.546	93.8	0.456	93.7
Joint model	Audio	0.482	95.2	0.639	95.6	0.660	92.9	0.469	94.0	0.549	94.1	0.454	93.7
HEXACO model	Visual	0.220	94.5	0.495	94.7	0.305	89.9	0.443	93.6	0.204	92.5	0.278	93.0
Joint model	Visual	0.214	94.5	0.502	94.8	0.320	90.4	0.454	93.7	0.198	92.8	0.290	93.3
HEXACO model	Audio, Visual	0.477	95.1	0.627	95.2	0.681	93.0	0.551	94.3	0.541	94.0	0.491	93.0
Joint model	Audio, Visual	0.480	95.2	0.635	95.4	0.691	92.9	0.568	94.2	0.547	94.0	0.504	93.8
HEXACO model	Audio, Visual, Text	0.492	94.6	0.651	95.3	0.693	93.1	0.570	93.6	0.559	94.0	0.594	94.4
Joint model	Audio, Visual, Text	0.504	95.2	0.645	95.6	0.707	93.2	0.576	94.3	0.579	94.2	0.608	94.4
<i>Human evaluation</i>		0.401	92.6	0.497	93.3	0.744	93.1	0.555	92.5	0.592	92.8	0.536	92.3

Table 3 shows the correlations between the Big Five and HEXACO for the human and automatic evaluation using the joint model with audio, visual, and text inputs. The results of the human evaluation show that the correlations between the Big Five and HEXACO were as expected for the characteristics considered highly correlated between the Big Five and HEXACO. *Honesty-Humanity* in the HEXACO did not correlate highly with any of the traits in the Big Five. These results indicate that our experimental setups using our newly annotated dataset were convincing for evaluating apparent Big Five and HEXACO traits. Next, the results of the automatic evaluation show that the correlations were higher not only for the characteristics considered highly correlated between the Big Five and HEXACO but also for other traits. This is because correlations between the Big Five and HEXACO were excessively captured during the model training. While we could achieve recognition performance equivalent to human performance, we could not reproduce the way humans perceive impressions. Bringing the correlations between the Big Five and HEXACO closer to human evaluation will be a future challenge.

V. CONCLUSION

We demonstrated the first investigation of automatically recognizing observer-perceiving HEXACO traits from multimodal human behavior. We also introduced a novel joint modeling method of Big Five and HEXACO to consider the relation-

TABLE IV
CORRELATION MATRIX BETWEEN BIG FIVE AND HEXACO

	Human evaluation					
	<i>H</i>	<i>E</i>	<i>X</i>	<i>A</i>	<i>C</i>	<i>O</i>
<i>O</i>	0.134	-0.155	0.479	0.378	0.539	0.797
<i>C</i>	0.432	0.066	0.170	0.464	0.837	0.518
<i>E</i>	-0.363	-0.301	0.937	0.114	-0.05	0.355
<i>A</i>	0.362	0.179	0.462	0.7621	0.430	0.558
<i>N</i>	0.078	-0.517	0.643	0.424	0.266	0.438
Automatic evaluation with the proposed joint model						
	<i>H</i>	<i>E</i>	<i>X</i>	<i>A</i>	<i>C</i>	<i>O</i>
<i>O</i>	0.299	-0.002	0.564	0.515	0.805	0.947
<i>C</i>	0.652	0.333	0.297	0.727	0.924	0.850
<i>E</i>	-0.553	-0.247	0.984	0.105	-0.07	0.363
<i>A</i>	0.500	0.554	0.529	0.921	0.567	0.7183
<i>N</i>	-0.302	-0.440	0.833	0.224	0.189	0.524

ships between them. The experimental results demonstrated the effectiveness of the proposed joint modeling approach, showing improved recognition performance for both the Big Five and HEXACO traits. Future work includes bringing the correlations between the Big Five and HEXACO closer to human evaluation.

REFERENCES

- [1] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, vol. 53, pp. 2313–2339, 2020.

- [2] X. Zhao, Z. Tang, and S. Zhang, "Deep personality trait recognition: A survey," *Frontiers in Psychology*, 2022.
- [3] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J. C. S. J. Júnior, M. Madadi, S. Ayache, E. Viegas, F. Gürpınar, A. S. Wicaksana, C. C. S. Liem, M. A. J. van Gerven, and R. van Lier, "Modeling, recognizing, and explaining apparent personality from videos," *IEEE Transactions on Affective Computing*, vol. 13, pp. 894–911, 2022.
- [4] W. Ilmini and T. Fernando, "Detection and explanation of apparent personality using deep learning: a short review of current approaches and future directions," *Computing*, vol. 106, pp. 275–294, 2024.
- [5] Y. Güçlütürk, U. Güçlü, M. A. J. van Gerven, and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition," *In Proc. European Conference on Computer Vision (ECCV) workshops*, pp. 349–358, 2016.
- [6] J. Gorbova, E. Avots, I. Lüsü, M. Fishel, S. Escalera, and G. Anbarjafari, "Integrating vision and language for first-impression personality analysis," *IEEE Transactions on Multimedia*, vol. 25, pp. 24–33, 2018.
- [7] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," *In Proc. Association for Computational Linguistics (ACL)*, pp. 606–611, 2018.
- [8] R. D. P. Principi, C. Palmero, J. C. S. J. Júnior, and S. Escalera, "On the effect of observed subject biases in apparent personality analysis from audio-visual signals," *IEEE Transactions on Affective Computing*, pp. 607–621, 2021.
- [9] S. Aslan, U. Gündükbay, and H. Dibeklioglu, "Multimodal assessment of apparent personality using feature attention and error consistency constraint," *Image and Vision Computing*, vol. 110, p. 104163, 2021.
- [10] L. R. Goldberg, "An alternative description of personality: the big-five factor structure," *Journal of personality and social psychology*, pp. 1216–1229, 1990.
- [11] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, pp. 175–115, 1992.
- [12] K. Lee and M. C. Ashton, "Psychometric properties of the HEXACO personality inventory," *Multivariate behavioral research*, vol. 39, pp. 329–358, 2004.
- [13] M. C. Ashton and K. Lee, "Empirical, theoretical, and practical advantages of the HEXACO model of personality structure," *Personality and Social Psychology Review*, vol. 11, pp. 150–166, 2007.
- [14] K. Lee and M. C. Ashton, "Getting mad and getting even: Agreeableness and honesty-humility as predictors of revenge intentions," *Personality and Individual Differences*, vol. 52, no. 5, pp. 596–600, 2012.
- [15] L. Leone, A. Chirumbolo, and M. Desimoni, "The impact of the HEXACO personality model in predicting socio-political attitudes: The moderating role of interest in politics," *Personality and Individual Differences*, vol. 52, no. 3, pp. 416–421, 2012.
- [16] J. L. Pletzer, M. Bentvelzen, J. K. Oostrom, and R. E. de Vries, "A meta-analysis of the relations between personality and workplace deviance: Big Five versus HEXACO," *Journal of Vocational Behavior*, vol. 112, pp. 369–383, 2019.
- [17] P. Sinha, L. Dey, P. Mitra, and A. Basu, "Mining HEXACO personality traits from enterprise social media," *In Proc. Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pp. 140–147, 2015.
- [18] M. C. Ashton and K. Lee, "Honesty-humility, the big five, and the five-factor model," *Journal of personality*, vol. 73, pp. 1321–1354, 2005.
- [19] R. E. de Vries and J.-L. van Gelder, "Tales of two self-control scales: Relations with five-factor and HEXACO traits," *Personality and Individual Differences*, vol. 54, pp. 756–760, 2013.
- [20] M. C. Howard and E. C. V. Zandt, "The discriminant validity of honesty-humility: A meta-analysis of the HEXACO, Big Five, and Dark Triad," *Journal of Research in Personality*, vol. 87, p. 103982, 2020.
- [21] R. Liao, S. Song, and H. Gunes, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 12 113–12 132, 2023.
- [22] R. Masumura, S. Orihashi, M. Ithori, T. Tanaka, N. Makishima, S. Suzuki, S. Mizuno, and N. Hojo, "Multimodal fine-grained apparent personality trait recognition: Joint modeling of big five and questionnaire item-level scores," *In Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1456–1464, 2025.
- [23] L. R. Goldberg, "The structure of phenotypic personality traits," *American Psychologist*, vol. 48, pp. 26–34, 1993.
- [24] M. T. Apple and P. Neff, "Using rasch measurement to validate the big five factor marker questionnaire for a japanese university population," *Journal of Applied Measurement*, vol. 13, pp. 1–21, 2012.
- [25] M. C. Ashton and K. Lee, "The HEXACO–60: A short measure of the major dimensions of personality," *Journal of Personality Assessment*, vol. 91, pp. 340–345, 2009.
- [26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *In Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 5998–6008, 2017.
- [27] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv:1904.07850*, 2019.
- [28] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5525–5533, 2016.
- [29] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," *In Proc. International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.
- [30] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 29, pp. 3451–3460, 2021.
- [31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *In Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.
- [32] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2584–2593, 2017.
- [33] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, pp. 18–31, 2019.
- [34] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *In Proc. International Conference on Learning Representations (ICLR)*, 2020.
- [35] V. Ponce-López, B. Chen, M. Olliu, C. A. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "ChaLearn LAP 2016: First round challenge on first impressions - dataset and results," *In Proc. European Conference on Computer Vision (ECCV)*, pp. 400–418, 2016.
- [36] H. J. Escalante, I. Guyon, S. Escalera, J. C. S. J. Júnior, M. Madadi, X. Baró, S. Ayache, E. Viegas, Y. Güçlütürk, U. Güçlü, M. A. J. van Gerven, and R. van Lier, "Design of an explainable machine learning challenge for video interviews," *In Proc. International Joint Conference on Neural Networks (IJCNN)*, pp. 3688–3695, 2017.