

Automatic Personality Recognition in Videos using Dynamic Networks and Rank Loss

Nethravathi Periyapatna Sathyanarayana, Karuna Pandith, Manjula Sanjay Koti, Rajermani Thinakaran

Faculty of Computer Science, Shree Devi Institute of Technology, Mangalore, India¹

Department of Information Science & Engineering, N.M.A.M. Institute of Technology, Nitte, Karnataka, India²

Faculty-Dept. of MCA, Dayananda Sagar Academy of Technology & Management, Bangalore, India³

Faculty of Data Science and Information Technology, INTI International University, Negeri Sembilan, Malaysia⁴

Abstract—There are a few difficulties with current automatic personality recognition technologies. Two of these are discussed in this article. They use of very brief video segments or individual frames to come to conclusion with personality factors rather than long-term behavior; and absence of techniques to record individuals' facial movements for personality recognition. To address these concerns, this work first offers a unique Rank Loss for self-regulated learning of facial movements that uses the innate time related development of facial movements in lieu of personality traits. Our method begins by training a basic U-net type system that can predict broad facial movements from a collection of unlabeled face recordings. The robust model is frozen subsequently, and a series of intermediary filters is added to the architecture. The self-regulated education is then restarted, but only with films tailored to the individual. As a result, the weights of the learnt filters are individual-specific, making it a useful tool for simulating individual facial dynamics. The weights of the learnt filters are then concatenated as an individual-specific representation, to forecast personality factors without the assistance of other components of the network. The proposed strategy is tested on ChaLearn personality dataset. We infer that the tasks performed by the individual in the video matter, merging or combined application of tasks achieves the high-rise precision. Also, multi-scale characteristics are better penetrating than single-scale dynamics, along with achieving impressive outcomes as process innovation in prediction of the personality factors scores through videos.

Keywords—Automatic personality recognition; facial movements; individual-specific representation; personality factors; convolutional neural networks

I. INTRODUCTION

Human personality is a unique combination of actions, thoughts, and affective patterns that develop across time and space as a result of biological and environmental influences [1] and can be expressed in the consistent patterning of affect and behaviours [2][3][4][5][6]. Understanding human behaviour, emotional processes, and physical conditions can all be aided by recognizing personality. There are two forms of personality that can be assessed: 1) self-reported personality, that is more than one observer's view of an individual based on different cues; 2) considered personality, that again is more than one observer's impression of an individual based on different cues [7] [8].

Trait-based personality models, such as the Eysenck personality inventory, five-factor model of personality [9][10] are widely used and primarily focus on analyzing characteristics

of personality that are reasonably constant over time yet differ between individuals. While using verbal behaviour descriptors-based questionnaires is a common method of assessing personality traits, earlier psychological studies, have commonly stated that the non-facial behaviours also accommodate essential indications to a human's indirect disposal and internal state. As a result, nonverbal face cues are included in most video-based automatic personality diagnosis approaches. They typically try to learn personality from a very short segment or a single frame, reusing personality labeling (video-level) as the labels for its integral image components as well as training the machine learning models grounded upon such small fragments. Personality cannot be deduced only from a frame/short segment. While other research built video-level descriptors by enumerating global features of all levels (frame or segment) custom-made descriptors, the resulting features lacked comprehensive temporal information, which is a crucial component of face behaviour [10][11][12] [13].

Our goal in this work is to find an individual facial movement describer for every person that is fairly constant over a period of time but different from that of others. First, we present a self-supervised learning strategy based on Bilen *et al.*'s dynamic image [14]. On the contrary, we suggest using a predicate task in a self-regulated scenario to create an animated image from a sole image.

Second, we propose a domain adaptation strategy that uses adaptation layers to include individual-specific data into the trained network (Adaptive layer). Third, we propose, in contrast to other approaches, using the weights of the adaptive layers as an inception for the following task, in this instance personality trait prediction.

The rest of the paper are arranged in the following subsections as follows: Section II - consists of an extensive literature survey; Section III - is the detailed explanation of the methodology used in the study; Section IV - is a detailed presentation of the results arrived at along with comparisons. Lastly, Section V conclude the paper by providing future directions.

II. LITERATURE SURVEY

Automatic personality analysis based on video is a multidisciplinary study topic that combines psychology results with cutting-edge machine learning techniques. A popular lexical technique for personality evaluation is the Five Factor

Model (FFM) [10] it's among the most popular personality tests. The five personality traits that make up the FFM model are conscientiousness, extraversion, and openness to experience, agreeableness, and neuroticism. Anxiety, rage, concern, humiliation, insecurity, and feelings are all linked to neuroticism (also known as emotional stability). It is also believed that this dimension has two sub-dimensions: ambition and sociability. Conscientiousness, which is associated with conformity, desire to succeed, dependability (structured and accountable), and volition (hardworking, achievement-oriented, enduring), has been widely regarded as the third dimension. Flexible, tolerant, good-natured, forgiving, polite, and soft-hearted people are related with agreeableness, also known as likability. Extraversion, in particular, is typically linked to a gregarious, aggressive, talkative, energetic, and friendly personality. Finally, openness is a measure of a person's cognitive interest, inventiveness, as well as need for newness and variation. People who have high scores on directness are more prone to drop concentration and participate in dangerous behaviour [15].

The FFM model's resilience has been experimentally demonstrated in a variety of situations, including several theoretical frameworks, numerous, evaluation from various sources, and utilizing a variety of instances. The FFM model is also used as a personality evaluation in this research. A person's implicit inclinations and internal moods are mirrored in their nonverbal expressive behaviour, according to converging data, in particular, looked into the link between bodily clues and an individual's true nature [16][17][18].

Keltner [19] discovered that individual factors are represented uniquely, visible behaviours which elicit retaliation in others, and that a small number of universal features may be used to explain human nature.

Furthermore, certain research found that when subjects were photographed in a natural stance with a natural facial expression, observers' judgements were almost always correct. A significant amount of instantaneous personality prediction techniques based on face have lately been brought out, that can approximately be classified into two classes: video-level feature-based techniques and frame-level feature-based methods, as many studies on psychology have put forward that personality factors are labelled by facial exhibits, and automated study of facial movements is highly precise [20][21]. For estimating personality traits, Ilmini and Fernando [22] presented an audio-visual Residual network. This architecture is made up of two streams: an audio stream and a visual, both were utilized to extract frame-level deep audio and visual traits, and they were merged at the fully connected layer to offer frame-wise projections. For withdrawing both auditory and visual frame-level data, Wei et al., [23] presented a Deep Bimodal Regression Network. They used global average pooling or global max pooling for visual information and created Descriptor Aggregation Networks (DAN) to combine factors from various layers of convolution. The final forecast is also arrived at by

fusing frame-level projections, as is the case with previous techniques.

Nevertheless, all of the techniques narrated above make use of video level labels to train models for frame/segment-wise feature extraction, implying that they are all based on the assumption that personality can be casted back by a single facial display or a very short-duration facial action, which leads to poorly posed machine learning problems. To overcome this assumption, Biel, Teijeiro-Mosquera and Gatica-Perez [24] and Teijeiro-Mosquera et al., [25] instigated approaches that generated statistical evidence of facial movements from each frame or from a short segment. The researchers then withdrew four video-level visual indications to reflect the existence, time span, as well as frequency of facial movements that were then catered into a regressor to predict personality. Okada, Aran and Gatica-Perez [26] retrieved various hand-crafted visual and aural components frame by frame as a time-series to present binary non-verbal behaviours.

To sum up, while the majority of existing techniques try to deliver a unified group of personality characteristic predictions for every video, mostly they assume that the video-level label may be regarded as the frame/short segment-level label. This method would not just result in a vague machine learning architecture, but also go against the concept of personality factors, which is that they are firm over a period of time yet vary among people. A few research dedicated to video-level feature extraction did not use this assumption, they largely relied on assembled mid-level signals [26], which may have missed many key spatial-temporal patterns or led to temporal information loss [27][28]. In order to overcome all of these issues, this research introduces a novel individual-specific feature extraction approach.

III. METHODOLOGY

The authors explored comparable feature representation (FR) for self-regulated learning of facial movements because our objective is to train a network which grasps facial movements. Instead of learning a unique feature representation, for each picture series (as put forward in [29]), the authors want to build a generic network that can prognosticate the FR for different face image series based on a single (central) image. This forces the network to learn the generic temporal evolution of faces in any brief series. The complete flow of the paper is shown in Fig. 1.

In this study, face images is given by $F_t \in \mathbb{R}^{m \times n}$, and let $F_{t-T}, F_{t-T+1}, \dots, F_{t-1}$ and $F_{t+1}, F_{t+2}, \dots, F_{t+T}$ be the frames related to a frame size of $2T + 1$ (window size), where F_t is the center. The authors used self-supervised learning for FR for every single input image F_t and it is given as $f(\cdot, \theta)$ network where θ represents parameters. The image frame considered in the study is represented as S_a , i.e., $S_a = F_a$. Each frame a has a static representation S_a , $a \in [t - T, t + T]$.

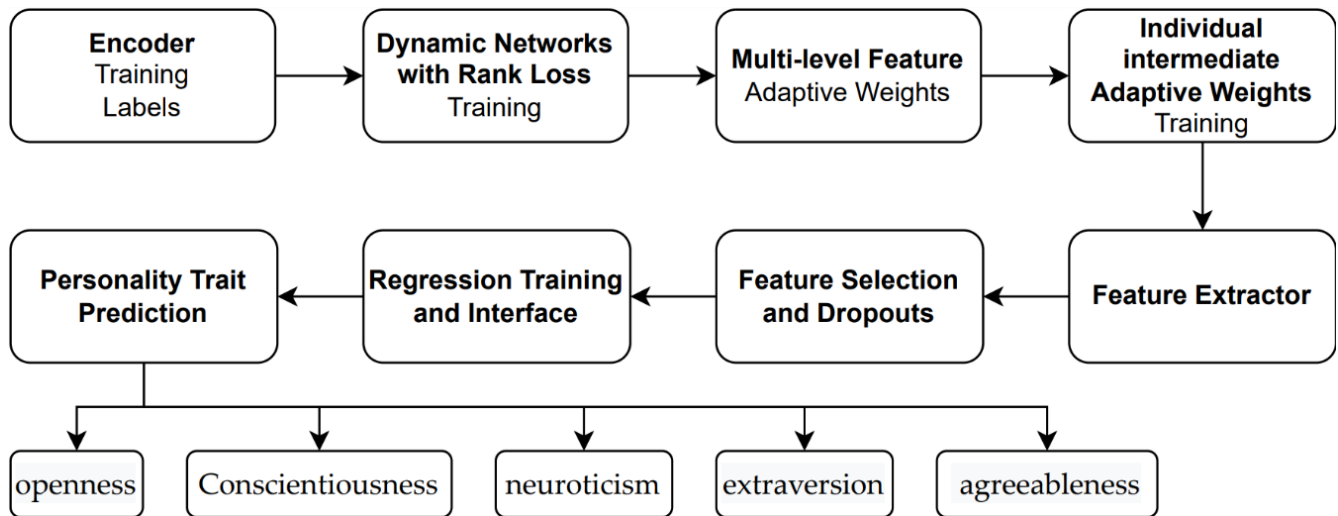


Fig. 1. An overview of the proposed system structure.

Every frame adjacent to F_t is used to extract the temporal information using FR. r_t replicates S_a in size that may rank previous and subsequent frames on the basis of their comparative temporal interval from F_t . The (Frobenius) inner product of the FR r_t is used to assign the score to every image and it is denoted by S_a , with a $[t T, t + T]$. The study proposes learning the FR by driving the architecture to create outcomes that instantly satisfy as in Eq. (1), i.e., the authors want the architecture to assist in determining the FR's parameters. So, the authors follow two steps to calculate the function S , which represents the score value for a given input image F_t . First step is to assign $r_t = f(F_t, \theta)$. The second step is to calculate the pairwise scores for all the frames within the $N = 2T + 1$ window as in Eq. (1).

$$\delta_{ab}(t) = X(r_t, S_a) - X(r_t, S_b) \quad (1)$$

The authors propose using a rank loss that exclusively differentiates negative scores, i.e. frames that will be ranked erroneously based on their scores. The study wish to emphasize that we are not backpropagation w.r.t. a pre-defined "ground-truth" r^*t . by employing the rank loss function. To put it another way, our strategy enables the network to be trained without any need to generate a target FR for each training image individually. In this sense, the network also helps to determine the FR's form.

1) *Structure of the network:* The ResNet network [30] is chosen as the network $f(., \theta)$ in this paper. (We call this network the Dynamic Network (DN) since it is a 5-layer encoder-decoder with numerous skip layers at various spatial resolutions). A 2-D convolution layer, an instance normalization layer, as well as a Leaky ReLU activation function make up each of the encoder's five blocks. There are an additional five blocks in the decoder. Each decoder block has a transposed convolution layer that duplicated the size of the input feature maps first. It is then catered into a Leaky ReLU and an instance normalization layer. Skip layers connect all five pairs of encoder-decoder layers. We propose a DN-adaptive layers network topology to infer individual-specific facial dynamics. To each DN skip layer (AL), we add a convolution

block comprising of a convolution layer having kernel size 1, an instance normalization layer, and a Leaky ReLU. For all individuals, the adaptive layer typologies are the same, i.e., five intermediate layers with a kernel size of 1, that contain DN's weights are frozen during person-specific training, and only the weights of the ALs that have been inserted have been modified.

2) *Training the models:* For personality recognition using artificial neural networks, a regression model is developed that takes the trained weights of individual-specific descriptors as input features for the ANN. It is made up of four hidden layers that are fully connected, and each layer has a dropout with a probability of 0.5. In order to minimize the issues of overfitting, dropouts are used in this study [31]. The output layer contains five nodes that correspond to the Big-Five personality qualities. This framework allows the model to grasp all five personality qualities at the same time.

We put forward to use multi-scale individual-specific descriptors for this work, to capture facial movements at a varied temporal scale, because the temporal scale of the person-specific representation is determined by the size of the time-window, while the appropriate timescale for personality analysis is still unknown. This can be accomplished by training numerous sets of ALs using a series of time-windows of varying lengths. For a certain person, the integration of these individual-specific characteristics represents face movements gathered at several temporal and spatial resolutions.

To sum up, the process for determining personality factors is as follows: we start with the DN network and then add the AL layers. The adaptation is then practiced for each subject, distinct layers are created, and the weights θ are used to reflect their features. The corresponding θ is then input into the ANN network. Ablation study contains more information goes over the specifics of network training implementation.

IV. RESULTS AND ANALYSIS

The ChaLearn [32] dataset was used to conduct apparent personality estimate studies, which employs the Big-Five personality factors as labels but normalizes their values to the

range of [0, 1]. The ChaLearn dataset includes 10,000 videos of 2,764 YouTube users conversing to the camera, organized into three subsets: training (6,000 videos), validation (2,000 videos), and test (2,000 videos). Fig. 2 illustrates sample dataset base on ChaLearn [32]. The videos are all about 15 seconds long and taken at 30 frames per second. The ChaLearn dataset videos

were re-sampled to 25 fps in our studies. Multiple human annotators used Amazon Mechanical Turk to obtain the personality factor labels in this database. To summarize, the Chalearn dataset differs from other datasets in the following ways: 1. kind of annotations; 2. number of films; 3. Time span of videos; and 4. recording settings.





















Agreeableness			
Authentic		Self-interested	
			
0.9230	0.9340	0.1098	0.0879
Conscientiousness			
Organized		Sloppy	
			
0.9708	0.9514	0.0873	0.1068
Extraversion			
Friendly		Reserved	
			
0.9158	0.9252	0.0521	0.0933
Neuroticism			
Comfortable		Uneasy	
			
0.9585	0.9791	0.1005	0.0872
Openness			
Imaginative		Practical	
			
0.9777	0.9582	0.0549	0.1113

Fig. 2. Sample images with range values of ChaLearn dataset [32].

Evaluation metrics: Two different metrics are used to evaluate the proposed method for personality analysis. The Root Mean Square Error (RMSE) is specified as in Eq. (2), and the Pearson Correlation Coefficient (PCC) is given as in Eq. (3).

$$RMSE = \sqrt{\frac{1}{n} \sum_{x=1}^n (m_x - n_x)^2} \quad (2)$$

$$PCC = \frac{cov(m,n)}{\sigma_m \sigma_n} \quad (3)$$

Refer as in Eq. (2), m_x is the x^{th} prediction in the prediction vector m , cov is the covariance, n_x is the equivalent ground-truth in the ground-truth vector n , while refer as in Eq. (3), σ_m and σ_n are the standard deviations of m and n , respectively.

Refer as in Eq. (4), the mean accuracy measurement ACC [32] is used to collate against past outcomes on the ChaLearn dataset (as the ACC is employed in the ChaLearn challenge) where V_y is the number of videos, and b_x and a_x are the predictions and labels, respectively.

$$ACC = 1 - \frac{1}{V_y} \sum_{x=1}^{V_t} |a_x - b_x| \quad (4)$$

Comparison with other approaches: The suggested method is compared to other known approaches for automated self-reported and evident personality assessment in this section. The 3ESA is used in the comparison. On the ChaLearn dataset, Table I shows the comparison between our approach to previous personality prediction algorithms that were based on videos. Our multi-scale model had the optimum average PCC and RMSE results, with the optimum PCC performance on many of the personality factors and the greatest RMSE results on all five personality factors. Furthermore, the PCC for the agreeableness characteristic was highest for the predictions from the DRN technique [33]. Our best system outperformed most current approaches in terms of ACC, producing the second-best detection results on three qualities and the greatest result on the agreeableness factor.

TABLE I. COMPARISON OF RESULTS WITH PROPOSED METHOD (CHALEARN DATASET)

Metric	Methods	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	Average
PCC	Güçlütürk et al., [34]	0.36	0.12	0.2	0.25	0.25	0.236
	Song et al., [35] - (BP)	0.37	0.3	0.34	0.36	0.32	0.338
	Wei et al., [33]	0.43	0.37	0.45	0.34	0.36	0.39
	Jaiswal, Song and Valstar [36]	0.3	0.05	0.22	0.22	0.2	0.198
	Song et al., [35] - (LF)	0.23	0.19	0.25	0.33	0.23	0.246
	Proposed (single scale)	0.4	0.28	0.36	0.38	0.39	0.362
	Proposed (multi-scale)	0.48	0.33	0.41	0.45	0.46	0.426
RMSE	Güçlütürk et al., [34]	0.15	0.14	0.15	0.15	0.14	0.146
	Song et al., [35] - (BP)	0.15	0.13	0.14	0.14	0.14	0.14
	Wei et al., [32]	0.14	0.12	0.13	0.14	0.13	0.132
	Jaiswal, Song and Valstar [36]	0.17	0.15	0.17	0.17	0.16	0.164
	Song et al., [35] - (LF)	0.2	0.15	0.16	0.14	0.15	0.16
	Proposed (single scale)	0.11	0.13	0.13	0.11	0.11	0.118
	Proposed (multi-scale)	0.02	0.1	0.1	0.12	0.11	0.09
ACC	Güçlütürk et al., [34]	0.9088	0.9097	0.9109	0.9085	0.9092	0.90942
	Song et al., [35] - (BP)	0.9165	0.9099	0.9178	0.9109	0.9117	0.91336
	Li et al., [37]	0.92	0.9176	0.9218	0.915	0.9191	0.9187
	Escalante et al., [38]	0.9019	0.9059	0.9073	0.8997	0.9045	0.90386
	Wei et al., [35]	0.9112	0.9135	0.9128	0.9098	0.9105	0.91156
	Bekhouché et al., [39]	0.9155	0.9103	0.9137	0.9082	0.91	0.91154
	Jaiswal, Song and Valstar [36]	0.8949	0.897	0.9001	0.8913	0.8975	0.89616
	Song et al., [36] - (LF)	0.886	0.8997	0.9061	0.9082	0.9035	0.9007
	Zhang, Peng and Winkler [28]	0.92	0.914	0.921	0.914	0.915	0.9168
	Proposed (single scale)	0.9213	0.9345	0.9111	0.9214	0.9088	0.91542
	Proposed (multi-scale)	0.9211	0.9411	0.9212	0.9311	0.9308	0.92706

1) *Ablation studies*: Encoder pre-training: Multiple settings for the time window hyperparameters, such as window length and stride, were used in the evaluation. $N = 2T + 1$ frames were made use in the temporal window around every input image (T preceding frames, T succeeding frames and the given frame). We used four distinct values of stride S to sample these N frames. Every image sequence used in the training has a range of N/S frames. For various combinations of $T = \{3, 5, 7, 9\}$ (i.e. $N = \{7, 11, 15, 19\}$) and stride $S = \{1, 2, 3, 4\}$, we test our DN's ranking capacity. The ranking is calculated using a window size of $N = 2T + 1 = 19$ frames, in the extreme instance, i.e. when $T = 9$ and $S = 4$, uniformly sampled from a sequence of $N * S = 76$ frames (more than three seconds). At test time, frames are chosen using the same sampling process as during the model's training. For all encoder-related experiments, the training and validation of DNs used the same settings as before.

In conclusion, the outcomes of this article reveal that pre-training using an emotion-guided encoder had no significant effect on ranking accuracy. The encoder's learning rate, in contrast, is critical in learning generic face dynamics. Lowering or freezing the encoder's learning rate led to reduced ranking accuracy. More crucially, when the initial learning rate was maintained, the emotion-guided encoder offered improved personality performance. The relatively high learning rate can force both the emotion-related and emotion-independent components of the encoder to learn frame-related temporal signals because we assumed that the majority of emotion-related dynamics are unimportant for frame ranking. However, the self-supervised training of the pretrained encoder may lead the taught model to retain some emotion data connected to personality.

V. CONCLUSIONS

A novel technique to automated personality analysis was developed in this research. The developed system begins with pretraining guided encoder, which is then used to train a DN architecture that learns broad short-term facial dynamics using the proposed rank loss. The training is self-supervised, which means that no manual annotations are required. Then, in DN, a convolution block is placed, which is learned for each individual independently, using the same self-supervised method. Consequently, the learnt weights adjust to the associated person's facial behavior, and we propose that these weights be used as the person-specific descriptor.

Combining existing face-based research with speech data to identify personality is a potential future project. We're primarily interested in determining the best network topology for self-supervised learning of personality characteristics using audio-visual and spoken information. Although our models are learnt to summarize face movements instead of character information, they may be used to other challenges where facial dynamics are significant since they are trained unsupervised and domain agnostic. The proposed technique might be extended to additional areas in the future. Furthermore, there arises a scarcity of large-scale multimedia self-reported personality databases that limits the use of deep learning algorithms in this area.

REFERENCES

- [1] S. Siyang, S. Jaiswal, E. Sanchez, G. Tzimiropoulos, L. Shen, and M. Valstar, "Self-supervised learning of person-specific facial dynamics for automatic personality recognition", *IEEE Transactions on Affective Computing*, 2021.
- [2] M. Sean, and A. Hay, "Digital and in-person interpersonal emotion regulation: the role of anxiety, depression, and stress", *Journal of Psychopathology and Behavioral Assessment*, 45, no. 1, 2023, pp. 256-263.
- [3] M. Jia, L. Yu, M. Jiao, P. Vijayarajnam, A. Sivarajah, Y. Hui, "An Exploratory Study on the Influencing Factors of Personal Development Motivation in Learners to Improve Quality Education", *Migration Letters*, 20, no. S3, 2023, pp. 85-92.
- [4] T. Rajermani, S. Chupra, and M. Batumalay, "Motivation assessment model for intelligent tutoring system based on mamdani inference system", *IAES International Journal of Artificial Intelligence*, 2023, 12, no. 1, pp. 189.
- [5] T. Rajermani, R. Ali, and W. Nor Al-Ashekin Wan Husin, "A Case Study of Undergraduate Students Computer Self-Efficacy from Rural Areas", *Int. J. Eng. Technol*, 2018, 7, pp. 270.
- [6] I. C. Huang, J. L. Lee, P. Ketheeswaran, C. M. Jones, D. A. Revicki, and A. W. Wu, "Does personality affect health-related quality of life? a systematic review," *PLoS one*, 2017, vol. 12, no. 3, pp. e0173806.
- [7] N. Ute, S. Straßburg, S. Sutharsan, C. Taube, M. Welsner, F. Stehling, and R. Hirtz, "How personality influences health outcomes and quality of life in adult patients with cystic fibrosis", *BMC Pulmonary Medicine*, 2023, 23, no. 1, pp. 190.
- [8] N. Mina, N. Moghadam Charkari, and M. Mansoorzadeh, "Automatic Facial Emotion Recognition Method Based on Eye Region Changes", *Journal of Information Systems and Telecommunication (JIST)*, 2016, 4, no. 4 221-231.
- [9] H. J. Eysenck and S. Eysenck, "The Eysenck personality inventory," 1965.
- [10] R. Lau Chloe, M. Bagby, B. G. Pollock, and L. Quilty, "Five-Factor Model and DSM-5 Alternative Model of Personality Disorder Profile Construction: Associations with Cognitive Ability and Clinical Symptoms", *Journal of Intelligence*, 2023, 11, no. 4, pp. 71.
- [11] K. Kenan, A. Kashevnik, A. Mayatin, and D. Zubok, "VPTD: Human Face Video Dataset for Personality Traits Detection", 2023, *Data* 8, no. 7 pp. 113.
- [12] N. Azamossadat, M. S. Moin, and A. Sharifi, "Facial Images Quality Assessment based on ISO/IEC 29591 Standard Compliance Estimation by HMAX Model", *Journal of Information Systems Telecommunication*, 2019, 7 pp. 225-237.
- [13] F. Hasan, and M. Hasheminejad, "Fast Automatic Face Recognition from Single Image per Person Using GAW-KNN", *Information Systems & Telecommunication*, 2014, pp.188.
- [14] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3034-3042.
- [15] B. Ambridge, "Psy-Q: You know your IQ-now test your psychological intelligence," *Profile Books*, 2014.
- [16] B. M. DePaulo, "Nonverbal behavior and self-presentation," *Psychological bulletin*, 1992, vol. 111, no. 2, pp. 203.
- [17] G. Zhiyun, W. Zhao, S. Liu, Z. Liu, C. Yang, and Y. Xu, "Facial emotion recognition in schizophrenia", *Frontiers in Psychiatry*, 2021, 12, pp. 633717.
- [18] L. P. Naumann, S. Vazire, P. J. Rentfrow, and S. D. Gosling, "Personality judgments based on physical appearance," *Personality and social psychology bulletin*, 2009, vol. 35, no. 12, pp. 1661-1671.
- [19] D. Keltner, "Facial expressions of emotion and personality", in *Handbook of emotion, adult development, and aging*. Elsevier, 1996, pp. 385-401.
- [20] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE transactions on affective computing*, 2017.

- [21] T. Murat, N. Kahraman, and C. Eroglu Erdem, "Face recognition: Past, present and future (a review)", *Digital Signal Processing*, 2020, 106, pp. 102809.
- [22] W. M. K. S. Ilmini, and T. G. I. Fernando, "Detection and explanation of apparent personality using deep learning: a short review of current approaches and future directions." *Computing*, 2023, pp.1-20.
- [23] X. S. Wei, C. L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, 2018, vol. 9, no. 3, pp. 303–315.
- [24] J. I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: predicting personality from facial expressions of emotion in online conversational video", in *Proceedings of the 14th ACM international conference on Multimodal interaction, ACM*, 2012, pp. 53–56.
- [25] L. Teijeiro-Mosquera, J. I. Biel, J. L. Alba-Castro, and D. Gatica- Perez, "What your face vlogs about: expressions of emotion and big-five traits impressions in youtube," *IEEE Transactions on Affective Computing*, 2015, vol. 6, no. 2, pp. 193–205.
- [26] S. Okada, O. Aran, and D. Gatica-Perez, "Personality trait classification via co-occurrent multiparty multimodal event discovery," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ACM*, 2015, pp. 15–22.
- [27] 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, vol. 2, pp. 606–611.
- [28] L. Zhang, S. Peng, and S. Winkler, "Persemon: A deep network for joint analysis of apparent personality, emotion and their relationship," *IEEE Transactions on Affective Computing*, 2019.
- [29] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action recognition with dynamic image networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [30] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognition*, 2019, vol. 90, pp. 119-133.
- [31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *In Thirty-first AAAI conference on artificial intelligence*, 2017.
- [32] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, IGuyon, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," *In European conference on computer vision*, 2016, October, pp. 400-418, Springer, Cham.
- [33] X. S. Wei, C. L. Zhang, H. Zhang, and J. Wu, "Deep bimodal regression of apparent personality traits from short video sequences," *IEEE Transactions on Affective Computing*, 2018, vol. 9, no. 3, pp. 303–315.
- [34] Y. Güçlütürk, U. Güçlü, M. A. van Gerven, and R. V. Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition", *In European conference on computer vision*, 2016, Oct 8, pp. 349-358, Springer, Cham.
- [35] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, 2020, vol. 13, Issue: 2 pp. 829-844.
- [36] S. Jaiswal, S. Song, and M. Valstar, "Automatic prediction of depression and anxiety from behaviour and personality attributes," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.
- [37] Y. Li, J. Wan, Q. Miao, S. Escalera, H. Fang, H. Chen, X. Qi, and G. Guo, "Cr-net: A deep classification-regression network for multimodal apparent personality analysis," *International Journal of Computer Vision*, 2020, pp. 1–18.
- [38] H. J., Escalante, H. Kaya, A. Ali Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J.C.J. Junior, M. Madadi, and S. Ayache, "Modeling, recognizing, and explaining apparent personality from videos", *IEEE Transactions on Affective Computing*, 2020, vol. 13, no. 2, pp. 894-911.
- [39] S. E. Bekhouche, F. Dornaika, A. Ouafi, and A. Taleb-Ahmed, "Personality traits and job candidate screening via analyzing facial videos," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE*, 2017, pp. 1660–1663.