A Novel Multi-Level Pyramid Co-Variance Operators for Estimation of Personality Traits and Job Screening Scores

Hichem Telli, Salim Sbaa, Salah Eddine Bekhouche, Fadi Dornaika, Abdelmalik Taleb-Ahmed, Miguel Bordallo López

IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 38, No. 3, June, 2021, pp. 539-546
Journal homepage: http://iieta.org/journals/ts

A Novel Multi-Level Pyramid Co-Variance Operators for Estimation of Personality Traits and Job Screening Scores

Hichem Telli[1*], Salim Sbaa[1], Salah Eddine Bekhouche[2], Fadi Dornaika[2,3], Abdelmalik Taleb-Ahmed[4], Miguel Bordallo López[5]

[1] Laboratory of LESIA, University of Biskra, Biskra 07000, Algeria
[2] Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Leioa 48940, Spain
[3] IKERBASQUE, Basque Foundation for Science, Bilbao 48009, Spain
[4] Université Polytechnique Hauts de France, Univ. Lille, CNRS, Centrale Lille, F-59313, Valenciennes, France
[5] VTT Technical Research Centre of Finland & University of Oulu, Oulu 90570, Finland

Corresponding Author Email: tellihicham@univ-biskra.dz

https://doi.org/10.18280/ts.380301

Received: 8 February 2021
Accepted: 10 June 2021

Keywords:
APA2016 dataset, Big-Five personality traits, job candidate screening, PML-COV descriptor, regression

ABSTRACT

Recently, automatic personality analysis is becoming an interesting topic for computer vision. Many attempts have been proposed to solve this problem using time-based sequence information. In this paper, we present a new framework for estimating the Big-Five personality traits and job candidate screening variable from video sequences. The framework consists of two parts: (1) the use of Pyramid Multi-level (PML) to extract raw facial textures at different scales and levels; (2) the extension of the Covariance Descriptor (COV) to fuse different local texture features of the face image such as Local Binary Patterns (LBP), Local Directional Pattern (LDP), Binarized Statistical Image Features (BSIF), and Local Phase Quantization (LPQ). Therefore, the COV descriptor uses the textures of PML face parts to generate rich low-level face features that are encoded using concatenation of all PML blocks in a feature vector. Finally, the entire video sequence is represented by aggregating these frame vectors and extracting the most relevant features. The exploratory results on the ChaLearn LAP APA2016 dataset compare well with state-of-the-art methods including deep learning-based methods.

1. INTRODUCTION

Automatic personality perception and synthesis explores how a machine might automatically recognize or synthesize human personality [1]. For a human, this is an instinctive task that assesses the personality of others at first glance, even without having any interaction with them. This human assessment usually occurs very quickly, in a small fraction of a second [2].

In the past decades, many studies have been conducted on personality traits and their classification. In this context, several models have been proposed, such as the Big-Five [3], BigTwo [4], or 16PF [5], among many others. The Big-Five (or Five-Factor Model) is a personality model widely used in the field of psychology. It characterizes an individual's personality based on five independent dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience [3].

Automatic personality analysis using computer vision is a relatively new research topic, where various applications can use personality analysis systems such as pre-screening interviews, personalized agents, criminal activities, and political ideology. The first competition in this field was the *ChaLearn Looking at People 2016 First impression challenge* [6]. It targeted researchers around the world to try to solve the problem of identifying these Big-Five personality traits from facial videos. Due to its success, another competition was proposed. The *ChaLearn Looking at People CVPR 2017 Challenge* came with an extension of the problem, namely adding a score for the screening attribute of the applicant to be estimated along with the Big-Five personality traits.

In this paper, we propose the use of a computationally efficient hand-crafted descriptor that can extract low-level facial features from video sequences. This descriptor naturally merges multiple local texture features using a Pyramid Multi-Level (PML) representation [7] and a Co-Variance Operator. It extracts and fuses information from multiple scales and face regions.

Inspired by our previous work on descriptors used to discriminate between classes [8, 9], the current study includes the following modifications: (i) improving the selection of low-level image descriptors that feed the Co-Variance Operator (COV), (ii) modification of the feature selection scheme to produce a real score value, and (iii) application of the descriptor to a regression task from facial videos.

The contributions of the paper can be summarized as follows:

1) A novel Pyramid Multi-Level Co-Variance Operator descriptor (PML-COV); a low computational cost descriptor that extends Co-Variance Operator to solve regression problems from videos.
2) The application of the novel descriptor to obtain state-of-the-art results in estimating personality traits and job screening scores, using benchmark datasets.

The remainder of the paper is organized as follows: Section 2 presents a summary of existing methods for estimating the

Big-Five personality traits and the interview score. an introduction of our proposed approach in section 3. The experimental results are given in section 4. In section 5, we present the conclusion and some perspectives.

2. RELATED WORK

In recent years, many experts have made some advances in the field of personality assessment by evaluating personality traits using visual information. The first survey on automatic personality detection, perception, and synthesis was presented by Vinciarelli and Mohammadi [1]. It summarizes the models based on features that most effectively predict measurable aspects in people's lives. In 2018, Escalante et al. [10] reviewed and investigated the mechanisms related to first impression analysis, and summarized the results of the CVPR 2017 Challenge, while the most recent review of previous image-based approaches to overt personality trait detection is presented by Jacques Junior et al. [11].

Approaches to the problem are numerous and diverse in nature. For example, Chen et al. [12] used a pairwise ranking approach to avoid calibration problems, and a total of 321,684 pairs were used on the ChaLearn Apparent Personality Trait Dataset. Based on Deep Learning, Ventura et al. [13] used the CNN architecture proposed by Zhang et al. [14], focusing only on the video modality to estimate the interview.

The addition of multiple sources of information has been shown to be a viable approach by merging audio and visual modalities. Zhang et al. [14] proposed a Deep Bimodal Regression (DBR) framework in which they modify the traditional convolutional neural networks to utilize important visual cues. While Subramaniam et al. [15] proposed two end-to-end trained deep learning models using facial images and audio features for first impression recognition. They trained their network with temporally ordered audio and novel stochastic visual features from a few frames, taking care to avoid overfitting.

The modalities can also be extended to the semantic domain by including text and speech information. Gorbova et al. [16] developed a multimodal method that uses audio, video, and NLP features as input to three separate LSTMs. The results of these LSTM networks are processed by a linear regressor and finally fed to an output layer. Kampman et al. [17] proposed a multimodal audio, text, and video architecture that uses a CNN model for each mode. The outputs of these CNNs are concatenated and passed to two fully connected layers to obtain the Big-Five personality features. Vo et al. [18] developed a multimodal framework that uses visual, auditory, and textual information using a cascade network built on advanced gradient boosting algorithms to generate the Big-Five personality features.

For job screening score detection, Güçlütürk et al. [19] proposed audio-visual deep learning-based architectures. They trained two deep residual networks on the ChaLearn dataset with 17-layers each for audio and visual streams, then an audio-visual stream of a fully connected layer outputting the Big-Five features. Finally, they used a linear regression model to explain the interview decision based on the predictions of these features. Kaya et al. [20] developed a multimodal approach that combines face and audio features and feeds them into modality-specific regressors to predict revealed personality features using an ensemble of decision trees. They also used another decision tree, combined with a rule-based algorithm, to explain interview decisions.

Several teams obtained competitive results with different strategies, the FDMB team used frame differences and Local Phase Quantization descriptors at several fixed image regions with the support vector regression (SVR) technique to predict the interview variable and the Big-Five traits [10]. While the ROCHCI team extracted a set of multimodal features; firstly using the SHORE library [21] to obtain visual information, and secondly using the audio signal to obtain pitch and intensity features. Finally, these features were combined, and a gradient boosting regression algorithm was applied to predict the personality traits and jobs screening score.

Hand-crafted descriptors are considered simple and suitable for real-time applications, as they can be easily deployed in low-cost hardware. However, they depend on perfect face alignment, so they are vulnerable to difficult face pose scenarios. On the other hand, deep learning-based approaches are very good at solving highly complex problems and can be easily applied to similar problems. However, they rely on very expensive hardware, and their training is time-consuming. Moreover, they mainly depend on data abundance and require careful selection of network design and hyperparameters.

3. PROPOSED APPROACH

In our work, six steps are performed to obtain the interview score, which are: (i) face preprocessing, where the detected faces are aligned and cropped, (ii) feature extraction, where a set of features for each face is extracted, (iii) video descriptor computation, where the features of each video are computed by taking the mean of all face feature vectors, (iv) feature selection, where a non-parametric feature selection method is applied to exclude possible irrelevant features, (v) personality traits estimation, and (vi) interview variable estimation. Figure 1 illustrates the general structure of our approach.

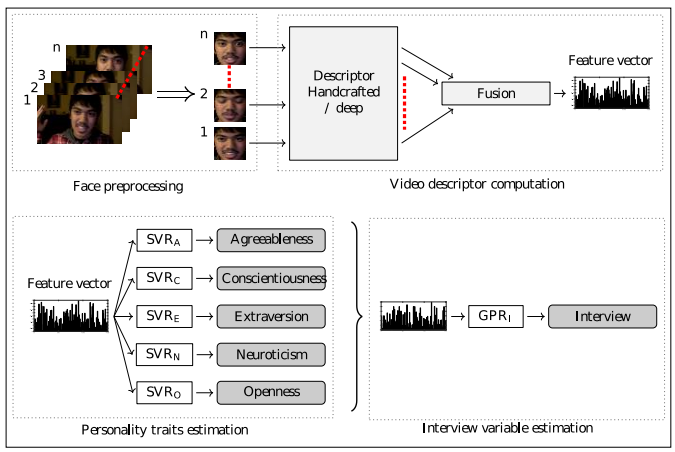


Figure 1. The general structure of the proposed approach

3.1 Face preprocessing

In the face preprocessing stage, we first detect the face region using the Viola-Jones [22] algorithm, and then we estimate the eyes' position based on Dlib [23] facial landmarks ($I_1 \ldots I_{68}$). In our case, we only used four of these points which represent the point features of each eye referenced by I_{37}, I_{40} for the left eye and I_{43}, I_{46} for the right eye. Afterward, we rectify the face pose based on these positions. To this end, we

applied a 2D similarity transform to align and crop the original grayscale face.

3.2 Feature extraction

In this work, we compared multiple kinds of hand-crafted features and deep features to determine which one is more suitable for this task, where the details of each kind of features are mentioned below.

3.2.1 Hand-crafted features

Hand-crafted descriptors are either simple or sophisticated algorithms that extract the features through the information in the image itself. In our work, we used five kinds of hand-crafted descriptors: Local Directional Pattern (LDP) [24], Local Binary Patterns (LBP) [25], Local Phase Quantization (LPQ) [26], Binarized Statistical Image Features (BSIF) [27], and Co-Variance Operator descriptor (COV) [28]. Knowing that we apply Pyramid multi-level (PML) representation, which was first introduced by Bekhouche et al. [7], to all of them. PML allows to extract multi-level multi-scale features, so that each face will be divided into $\sum_{i=1}^{l} i^2 = l \times (l+1) \times (2 \times l + 1)/6$ sub-blocks called regions. We applied one of these descriptors at each time on all blocks generated by the PML face representation and the concatenation of these block-features produces the current face feature vector as depicted in Figure 2. In this work, we used PML level l=7, which leads to $B = \big(l \times (l+1) \times (2 \times l + 1)\big)/6 = 140$ sub-blocks.

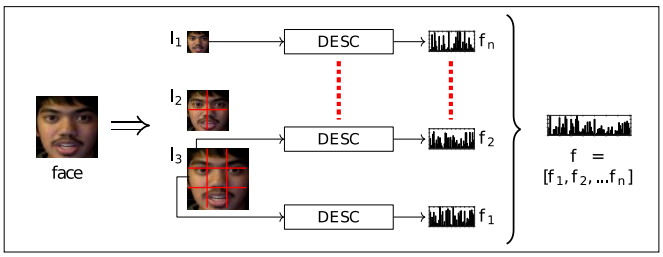


Figure 2. Feature extraction using PML-DESC descriptor

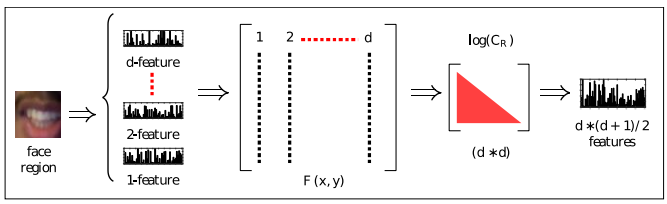


Figure 3. Covariance descriptor

Local Binary Patterns (LBP) [25], Local Phase Quantizations (LPQ) [26], Local Directional Pattern (LDP) [24], and Binarized Statistical Image Features (BSIF) [27] are a set of simple yet very efficient texture operators which labels the regions of an image by filtering the neighborhood of each pixel and considering the result as a binary number. The filters are handmade and encode intensity differences (LBP), locally computed phases (LPQ), directional edges (LDP), or statistical features derived from natural images (BSIF).

These set of features can be utilized to compose a descriptor using Co-Variance Operators. The Covariance descriptor (COV) was proposed by Tuzel et al. [28] as a region descriptor that could be used in object detection and texture classification problems. It takes advantage of the information provided by covariance matrices, which provide a natural way of fusing multiple features while keeping a low-dimensionality space due to its symmetry. Covariance matrices have only $d \times (d+1)/2$different values.

The COV descriptor (see Figure 3) is computed as follows: Let I denote $M \times N$ intensity image, and F be the $M \times N \times d$ dimensional feature image extracted from I, which contain a collection of image features such as horizontal coordinate, vertical coordinate, intensity, image gradient, or any image feature array. This dimensional feature image can be written as $F(x,y) = \phi(I,x,y)$, where ϕ is the mapping function of each feature image in this collection. For a given region $R \subset F$ containing s points, let $\{v_i\}_{i=1..s}$ be the d-dimensional feature points inside R. The region R is described by $d \times d$ covariance matrix of the feature points (See Eq. (1)). This region R can be characterized by $\log(C_R)$, where $\log(C_R)$ is the matrix logarithm of the square matrix C_R.

$$C_R = \frac{1}{s-1}\sum_{i=0}^{s}(v_i - \mu)(v_i - \mu)^T \tag{1}$$

where, μ is the mean of the points.

In this work, for the covariance descriptor, we define the mapping of d=19 channels as follows:

$$\begin{aligned} F(x,y) = [&x, y, I, Ix, Iy, Ixx, Iyy, \\ &LDP(k=3), LDP(k=5), \\ &LDP(k=7), BSIF(f=9\times9), \\ &BSIF(f=11\times11), BSIF(f=13\times13), \\ &LPQ(ws=7), LPQ(ws=9), \\ &LPQ(ws=11), LBP(r=1,n=8), \\ &LBP(r=2,n=8), LBP(r=2,n=16)]^T \end{aligned}$$

where, each descriptor in the dimensional feature image has been fed by a grayscale image which results to 19 2D arrays as follows: x and y are the pixel location, I is the intensity, I_x,I_y,I_{xx},I_{yy} are the first and second spatial intensity derivatives, $LDP(k)$ is LDP image obtained for a given k={3,5,7} most prominent directions, $BSIF(f)$ is BSIF image obtained for a given texture filter of size f={9×9,11×11,13×13} and 8bit length, $LPQ(ws)$ is LPQ image obtained for a given window size ws={7,9,11}, and finally $LBP(r,n)$ is LBP image obtained for a given radius r={1,2} and number of neighboring points n={8, 16}, since the number of channels used is 19 then the COV descriptor for each region is described by $D=d \times (d+1)/2=190$ features. And the total number of features for the whole image is $B \times D=140 \times 190$, where B is the total number of blocks, D is the image descriptor size in each block.

3.2.2 Deep features

Deep features are generally extracted from one of the last layers of a convolutional neural network (CNN). In our work, we used the VGG16 architecture [29] trained on VGGFace dataset [30], ResNet-50 [31] trained on VGGFace2 dataset [32], SE-ResNet-50 [33] trained on VGGFace2 dataset, Additive Angular Margin Loss (ArcFace) [34] based on SE-ResNet50 and trained on MS-Celeb-1M [35] dataset and MobileFacenet [36] which also trained on MS-Celeb-1M dataset.

VGG16 is a large CNN network with 16 learnable layers which has about 138 million parameters where we extract the deep features from the FC7 linear layer that produces 4096

features. ResNet-50 is a convolutional neural network based on a residual learning framework, where layers within a network are reformulated to learn a residual mapping rather than the desired unknown mapping between the inputs and outputs. ResNet-50 produces 2048 features from the global average pooling layer. SE-ResNet-50 has a similar architecture to ResNet50 and it produces the same number of features, however, it uses Squeeze and Excitation modules instead of residual blocks. ArcFace is a convolutional neural network that benefits from ResNet residual blocks and uses discriminative loss functions that directly optimize the parameters. The features of this network are extracted from the last linear layer in which 512 features are produced. MobileFacenet is a small convolutional neural network with less than 1 million parameters, its architecture is derived from MobileNetV2 [37] and uses ArcFace loss function. It produces the same number of features as ArcFace.

3.3 Video descriptor computation

To obtain the spatiotemporal feature vector for each video in the dataset, we compute the mean of all feature vectors as seen in Figure 1. Then, we apply L_2 feature normalization. This normalized mean vector will be used to represent the information of a whole video sequence. The mean statistical descriptor has been chosen based on the experiments shown in Table 2.

3.4 Feature selection

For feature selection, we used Neighborhood Component Analysis (NCA), which is a non-parametric learning method for estimating the feature weights. NCA sorts the features according to their relevance by performing feature ranking with regularization to learn feature weights, minimizing an objective function that measures the average leave-one-out classification or regression loss over the training data [38].

Based on NCA results we aimed to identify the best subset of features. NCA ranked features were used in the selection of the most relevant features for each personality trait separately. Thus, for each personality trait, we select the best features subset based on these ranked feature weights. The selected subset of features is determined by taking all features with high weights until reaching the mode of the histogram of the weights. Table 1 shows the averaged amounts (in %) of the selected features subset used for each method. The average was computed over the five traits.

Table 1. Amounts of selected features in % (averaged over the five regressors) using Neighborhood Component Analysis (NCA)

	VGG-FACE	ResNet-50	SE-ResNet-50	Mobilefacenet	Arcface ir-se-50
Mean	55.32	56.42	59.72	40.94	56.95
	PML-LDP	PML-BSIF	PML-LPQ	PML-LBP	PML-COV
Mean	99.95	98.80	90.35	86.11	70.86

3.5 Personality traits estimation

In the fifth stage, in order to estimate the scores of the Big-Five personality traits, we fed the five features' subsets, which we got after feature selection to five Support Vector Regressors (SVRs), one for each. These SVRs use hyperparameter optimization to improve the final performance and standardize the features using their corresponding weighted means and weighted standard deviations.

3.6 Interview estimation

The estimated five scores are then considered as a new feature vector, which we fed to a Gaussian Process Regression (GPR) scheme [39] to estimate the interview score. This GPR also uses hyperparameter optimization and standardizes the features to improve the interview score. The reason for choosing GPR instead of SVR for interview estimation is due to its high accuracy on very low dimensional data, which has been experimentally found.

4. EXPERIMENTS

4.1 Experimental settings

In this work, we used the dataset ChaLearn LAP 2016 APA [6]. This dataset consists of 10,000 short clip video sequences with an average duration of 15 seconds each, the resolution of the videos varies between 682x406 and 720x1280, and the number of frames varies between [49-456]. These video sequences were retrieved from YouTube and include more than 3,000 subjects. The subjects spoke English in front of a camera. The subjects depicted in the clips have different ages, genders, nationalities, and ethnicities.

The competition consisted of two phases, a validation phase and a testing phase. In the first phase, participants had access to 6,000 labeled video sequences, representing 60% of the dataset as a training set, and 2,000 unlabeled videos, representing 20% as a validation set. In the second phase, participants had access to the labeling of the previous validation set, and access to an additional 2,000 unlabeled videos as a test set.

For each video in the dataset, the ground truth labels for the Big-Five personality traits and interview variable were given by real values that fit the range [0, 1]. We used mean accuracy to evaluate performance for each personality trait and the interview variable. This indicator was used in the previous challenge [6]. This is given by P=100×(1-MAE), where MAE is the mean absolute error over the tested video sequences.

Table 2. Comparison of performances of different statistical descriptors for video descriptor computation based on interview score

Statistical descriptors	Validation	Test
Crest factor	91.57	91.55
Crest pulse	88.31	88.17
Kurtosis	90.86	90.88
Mean	92.14	92.11
Peaks	91.76	91.71
RMS	92.14	92.04
Shape factor	88.30	88.18
Skewness	91.79	91.69
Variance	92.01	91.89
Mean & RMS	92.14	92.07

Table 3. Comparison of performances of the proposed PML-COV descriptor with other hand-crafted and deep descriptors

	Method	AGRE	CONS	EXTR	NEUR	OPEN	MEAN	INTER
Deep	VGG-FACE	91.04	91.54	91.16	90.94	90.94	91.12	91.86
	ResNet-50	90.89	91.94	91.71	91.11	91.09	91.35	92.04
	SE-ResNet-50	90.92	91.90	91.63	91.01	90.80	91.25	91.93
	Mobilefacenet	91.10	91.25	91.21	90.31	91.10	90.99	91.52
	Arcface	91.09	91.52	91.49	90.72	91.01	91.17	91.88
Hand-crafted	PML-LDP	90.89	91.22	91.05	90.28	90.76	90.84	91.44
	PML-LBP	91.22	91.71	91.87	90.99	91.29	91.42	91.89
	PML-LPQ	91.30	91.93	91.95	91.06	91.35	91.52	91.92
	PML-BSIF	91.22	91.87	91.82	91.11	91.29	91.46	91.94
	PML-COV	91.32	92.03	91.91	91.06	91.31	91.53	92.11

To identify the optimal set of features with PML-COV, we apply other statistical descriptors besides the mean to all feature vectors of each video, such as the standard deviation, skewness, and kurtosis. Table 2 shows the results of these statistical descriptors and the combination of the top two descriptors (Mean & RMS) in terms of accuracy on both validation and test subsets.

4.2 Results and discussion

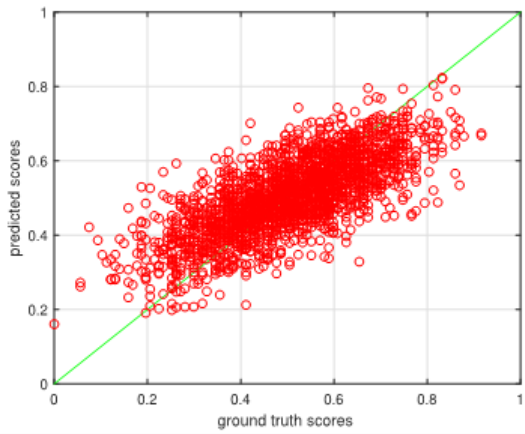
In our experiments, we performed two types of video feature extraction. First, using five different hand-crafted descriptors with the same PML face image representation. Second, using five different CNN models. This variation of methods is performed to determine which method is more suitable for our case. Table 3 summarizes the performance obtained on the test set. As shown, the PML-COV descriptor outperforms the other methods due to its ability to encode low-level facial features and its ability to fuse multiple known image texture descriptors. PML-COV is very fast in extracting the features of the video sequence, it requires less time and effort. The CNN-based method needs to be tuned to a specific or similar task as the target problem before extracting rich features from it, this training is time-consuming due to heavy operations. Table 4 summarizes the achieved performance of PML-COV descriptor on validation and test subsets.

Table 4. PML-COV results for validation and test subsets

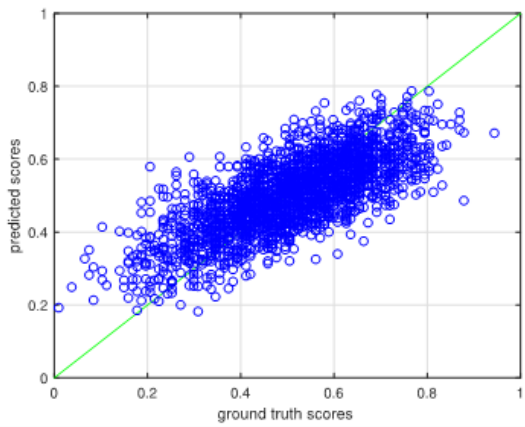
Trait	Validation	Test
AGRE	91.67	91.32
CONS	91.93	92.03
EXTR	91.81	91.91
NEUR	91.34	91.06
OPEN	91.47	91.31
INTER	92.14	92.11

Figure 4 shows the performance of the proposed approach. It illustrates the correlation between the ground-truth and predicted scores (validation and test sets) for the interview variable when the PML-COV descriptor is used. This correlation can be measured by a single measure, which is defined by Pearson correlation coefficient (PC) [40]. PC measures the linear correlation between ground-truth scores and estimated scores. In our experiment, the PC was 0.7297 for the validation set, and 0.7335 for the test sets. This indicates a good linear correlation between prediction and ground truth, as a perfect prediction would have a PC equal to one.

Table 5 summarizes the results obtained by state-of-the-art methods and schemes when they used the same dataset. PML-COV outperformed the other methods, although PML-COV use only visual information, while most of the other competing methods used multimodal feature fusion. In addition, many of these approaches relied on deep learning, a computationally very expensive and time-consuming method. The CPU time associated with each stage is given in Table 6, for the computation time, both the total test (2000 videos) and the average (1 video) are given. The test is performed on a custom workstation (Intel Xeon Processor E5-2658 v3, 30M Cache, 2.20 GHz, 64GB RAM, Windows 10).



(a) Validation



(b) Test

Figure 4. Correlations between true interview and estimated interview by the PML-COV descriptor

Table 5. A comparison of the proposed approach with other automatic personality estimation approaches

Approach	Deep Learning	AGRE	CONS	EXTR	NEUR	OPEN	MEAN	INTER
DAN+ [13]	YES	91.20	91.40	91.50	90.70	91.00	91.16	-
DRN-Baseline [41]	YES	91.02	91.38	91.07	90.89	91.11	91.09	-
Evolgen [15]	YES	91.19	91.19	91.50	90.99	91.17	91.21	-
NJU-LAMDA [14]	YES	91.26	91.66	91.33	91.00	91.23	91.30	-
FDMB [10]	NO	89.10	86.59	87.88	86.32	87.47	87.47	87.21
ROCHCI [10]	NO	90.32	89.49	90.26	90.11	90.47	90.13	90.18
PML [42]	NO	91.03	91.37	91.55	90.82	91.00	91.15	91.57
Baseline [19]	YES	91.12	91.52	91.12	91.03	91.11	91.18	91.62
BU-NKU [20]	YES	91.37	91.97	92.12	91.46	91.70	91.72	92.09
PML-COV (ours)	NO	91.32	92.03	91.91	91.06	91.31	91.53	92.11

Table 6. CPU time (seconds) of the different stages of our proposed framework

Stage	Task	Testing time (2000 videos)	Average (1 video)
Prepossessing	Detection and landmarks	8022.0	4.0110
	Alignment and crop	3302.1	1.6511
Feature extraction	Video descriptor computation	57219.0	28.6095
Estimation	BIG-5	180.6556	0.0903
	Interview	4.8262	0.0024
	Total	68728.5818	34.3643

5. CONCLUSIONS

In computer vision, apparent personality analysis from videos is a challenging problem. In this paper, we present a novel framework for evaluating Big-Five personality traits and screening attributes of job candidates from facial videos, and demonstrate its performance in solving the regression problem by comparing it with other state-of-the-art methods. This framework is based on two main components: (i) (PML) face representation to generate multiscale multiblocks in which face part properties are efficiently encoded, and (ii) the Covariance descriptor (COV) for face image analysis, which is capable of extracting rich and discriminative low-level face features. The proposed approach achieves high accuracy that outperforms the state-of-the-art results including deep CNNs. We also conducted an extensive experiment to compare hand-crafted features and deep features. The current work is limited to the video modality, but incorporating other modalities such as speech and text could improve the performance of the system. In future work, we could apply this framework to other spatiotemporal face analysis problems, such as pain assessment, disguised face identification, and drowsy driving detection.

ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors declare no conflict of interest.

REFERENCES

[1] Vinciarelli, A., Mohammadi, G. (2014). A survey of personality computing. IEEE Transactions on Affective Computing, 5(3): 273-291. https://doi.org/10.1109/TAFFC.2014.2330816

[2] Willis, J., Todorov, A. (2006). First impressions: Making up your mind after a 100-Ms exposure to a face. Psychological Science, 17(7): 592-598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

[3] McCrae, R.R., John, O.P. (1992). An introduction to the five-factor model and its applications. Journal of Personality, 60(2): 175-215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

[4] Abele, A.E., Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. Journal of Personality and Social Psychology, 93(5): 751. https://doi.org/10.1037/0022-3514.93.5.751

[5] Qin, R., Gao, W., Xu, H., Hu, Z. (2016). Modern physiognomy: An investigation on predicting personality traits and intelligence from the human face. arXiv preprint arXiv:1604.07499.

[6] Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions-dataset and results. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9915. Springer, Cham. https://doi.org/10.1007/978-3-319-49409-8_32

[7] Bekhouche, S.E., Ouafi, A., Dornaika, F., Taleb-Ahmed, A., Hadid, A. (2017). Pyramid multi-level features for facial demographic estimation. Expert Systems with Applications, 80: 297-310. https://doi.org/10.1016/j.eswa.2017.03.030

[8] Moujahid, A., Dornaika, F. (2018). A pyramid multi-level face descriptor: Application to kinship verification. Multimedia Tools and Applications, 78: 9335-9354. https://doi.org/10.1007/s11042-018-6517-0

[9] Moujahid, A., Dornaika, F. (2019). Multi-scale multi-block covariance descriptor with feature selection. Neural Computing and Applications, 32: 6283-6294. https://doi.org/10.1007/s00521-019-04135-7

[10] Escalante, H.J., Kaya, H., Salah, A.A., Escalera, S., Gucluturk, Y., Guclu, U. (2018). Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos. arXiv preprint arXiv:1802.00745.

[11] Jacques Junior, J.C.S., Güçlütürk, Y., Pérez, M., Güçlü, U., Andujar, C. (2018). First impressions: A survey on computer vision-based apparent personality trait analysis.

arXiv preprint arXiv:1804.08046.
[12] Chen, B., Escalera, S., Guyon, I., Ponce-López, V., Shah, N., Simón, M.O. (2016). Overcoming calibration problems in pattern labeling with pairwise ratings: application to personality traits. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9915. Springer, Cham. https://doi.org/10.1007/978-3-319-49409-8_33
[13] Ventura, C., Masip, D., Lapedriza, A. (2017). Interpreting CNN models for apparent personality trait regression. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1705-1713. https://doi.org/10.1109/CVPRW.2017.217
[14] Zhang, C.L., Zhang, H., Wei, X.S., Wu, J. (2016). Deep bimodal regression for apparent personality analysis. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9915. Springer, Cham. https://doi.org/10.1007/978-3-319-49409-8_25
[15] Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., Mittal, A. (2016). Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9915. Springer, Cham. https://doi.org/10.1007/978-3-319-49409-8_27
[16] Gorbova, J., Avots, E., Lüsi, I., Fishel, M., Escalera, S., Anbarjafari, G. (2018). Integrating vision and language for first-impression personality analysis. IEEE MultiMedia, 25(2): 24-33. https://doi.org/10.1109/MMUL.2018.023121162
[17] Kampman, O., Barezi, E.J., Bertero, D., Fung, P. (2018). Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. arXiv preprint arXiv:1805.00705.
[18] Vo, N.N., Liu, S., He, X., Xu, G. (2018). Multimodal mixture density boosting network for personality mining. In: Phung D., Tseng V., Webb G., Ho B., Ganji M., Rashidi L. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2018. Lecture Notes in Computer Science, vol 10937. Springer, Cham. https://doi.org/10.1007/978-3-319-93034-3_51
[19] Güçlütürk, Y., Güçlü, U., Baro, X., Escalante, H.J., Guyon, I., Escalera, S., Van Lier, R. (2017). Multimodal first impression analysis with deep residual networks. IEEE Transactions on Affective Computing, 9(3): 316-329. https://doi.org/10.1109/TAFFC.2017.2751469
[20] Kaya, H., Gürpinar, F., Salah, A.A. (2017). Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1651-1659. https://doi.org/10.1109/CVPRW.2017.210
[21] Ruf, T., Ernst, A., Küblbeck, C. (2011). Face detection with the sophisticated high-speed object recognition engine (SHORE). In: Heuberger A., Elst G., Hanke R. (eds) Microelectronic Systems. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23071-4_23
[22] Viola, P., Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. I–I. https://doi.org/10.1109/CVPR.2001.990517
[23] King, D.E. (2009). Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 10: 1755-1758. http://jmlr.org/papers/v10/king09a.html
[24] Jabid, T., Kabir, M.H., Chae, O. (2010). Local directional pattern (LDP)–A robust image descriptor for object recognition. 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 482-487. https://doi.org/10.1109/AVSS.2010.17
[25] Ahonen, T., Hadid, A., Pietikainen, M. (2006). Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 28: 2037-2041. https://doi.org/10.1109/TPAMI.2006.244
[26] Ojansivu, V., Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. International Conference on Image and Signal Processing, pp. 236-243. https://doi.org/10.1007/978-3-540-69905-7_27
[27] Kannala, J., Rahtu, E. (2012). Bsif: Binarized statistical image features. Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 1363-1366. https://ieeexplore.ieee.org/document/6460393
[28] Tuzel, O., Porikli, F.M., Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In: Leonardis A., Bischof H., Pinz A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3952. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11744047_45
[29] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[30] Parkhi, O.M., Vedaldi, A., Zisserman, A. (2015). Deep face recognition. Dans X. X. Jones., G. K. Tam (Éd.), Proceedings of the British Machine Vision Conference (BMVC) 41.1-41.12. BMVA Press. https://doi.org/10.5244/C.29.41
[31] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. https://doi.org/10.1109/CVPR.2016.90
[32] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67-74. https://doi.org/10.1109/FG.2018.00020
[33] Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132-7141. https://doi.org/10.1109/CVPR.2018.00745
[34] Deng, J., Guo, J., Zafeiriou, S. (2018). ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4685-4694. https://doi.org/10.1109/CVPR.2019.00482
[35] Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9907. Springer, Cham. https://doi.org/10.1007/978-3-319-46487-9_6
[36] Chen, S., Liu, Y., Gao, X., Han, Z. (2018). Mobilefacenets: Efficient CNNs for accurate real-time

face verification on mobile devices. In: Zhou J. et al. (eds) Biometric Recognition. CCBR 2018. Lecture Notes in Computer Science, vol 10996. Springer, Cham. https://doi.org/10.1007/978-3-319-97909-0_46
[37] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474
[38] Yang, W., Wang, K., Zuo, W. (2012). Neighborhood component feature selection for high-dimensional data. Journal of Computers, 7: 161. https://doi.org/10.4304/jcp.7.1.161-168
[39] Rasmussen, C.E., Williams, C.K. (2006). Gaussian Processes for Machine Learning. MIT Press. Récupéré sur https://books.google.dz/books?id=GhoSngEACAAJ.
[40] Olkin, I., Pratt, J.W. (1958). Unbiased estimation of certain correlation coefficients. The Annals of Mathematical Statistics, 29(1): 201-211. https://doi.org/10.1214/aoms/1177706717
[41] Güçlütürk, Y., Güçlü, U., van Gerven, M.A., van Lier, R. (2016). Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In: Hua G., Jégou H. (eds) Computer Vision – ECCV 2016 Workshops. ECCV 2016. Lecture Notes in Computer Science, vol 9915. Springer, Cham. https://doi.org/10.1007/978-3-319-49409-8_28
[42] Bekhouche, S.E., Dornaika, F., Ouafi, A., Taleb-Ahmed, A. (2017). Personality traits and job candidate screening via analyzing facial videos. 2017 IEEE Conference on Computer Vision and Pattern DSDDS Recognition Workshops (CVPRW), pp. 1660-1663. https://doi.org/10.1109/CVPRW.2017.211