Personality Traits Classification Using Deep Visual Activity-based Nonverbal Features of Key-Dynamic Images

Cigdem Beyan, *Member, IEEE,* Andrea Zunino, *Member, IEEE,* Muhammad Shahid, and Vittorio Murino, *Senior Member, IEEE*

Abstract—This paper addresses nonverbal behavior analysis for the classification of perceived personality traits using novel deep visual activity (VA)-based features extracted only from key-dynamic images. Dynamic images represent short-term VA. Key-dynamic images carry more discriminative information i.e., nonverbal features (NFs) extracted from them contribute to the classification more than NFs extracted from other dynamic images. Dynamic image construction, learning long-term VA with CNN+LSTM, and detecting spatio-temporal saliency are applied to determine key-dynamic images. Once VA-based NFs are extracted, they are encoded using covariance, and resulting representation is used for classification. This method was evaluated on two datasets: small group meetings and vlogs. For the first dataset, proposed method outperforms not only the state-of-the-art VA-based methods but also multi-modal approaches for all personality traits. For extraversion classification, it performs better than *i)* the most popular key-frames selection algorithm, *ii)* random and uniform dynamic image selection, and *iii)* NFs extracted from all dynamic images. Furthermore, the ablation study proves the superiority of proposed method. For the further dataset, it performs as well as the state-of-the-art visual-NFs on average, while showing improved performance for agreeableness classification. Proposed method can be adapted to any application based on nonverbal behavior analysis, thanks to being data-driven.

Index Terms—nonverbal behavior, visual activity, dynamic image, deep neural networks, long short-term memory, spatio-temporal saliency, key-frame, personality traits classification.

1 Introduction

Automatic nonverbal behavior analysis has been applied for various applications such as modeling dominance effects [1], [2], role recognition [3], [4], [5], [6], [7], emergent leader detection [8], [9], leadership style prediction [10], and personality traits classification [11], [12]. Among various nonverbal features (NFs), which are typically extracted from audio, video or by their fusion, the importance of visual activity (VA)-based NFs has been shown many times e.g., in [13], [14], [15], [16].

There are diverse way to detect the VA of a person. For example, in [9], [13], [14], [15] optical flow was used to detect the head activity, image differencing was applied to detect body activity and weighted motion energy image [17] was introduced to extract the motion in head and body together. As an alternative, 2-dimensional (2D) body pose representing VA was presented in [18]. Recently, to summarize the spatio-temporal content of a video in a single image, Bilen et al. [19] proposed dynamic image representation, which achieved significant results for activity recognition. That representation has never been used for nonverbal behavior analysis, and in this study, we adapt that method [19] to extract short-term VA.

Deep learning models, such as Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) have demonstrated the state-of-the-art results for action recognition and localization [19], [20], [21]. Their improvement can be exploited for VA-based nonverbal behavior analysis as well. Recently, [16] proposed deep VA-based NFs extracted from a hybrid CNN model [22] and applied them for emergent leader detection, leadership style prediction and low/high extraversion classification in small group meetings. Even though, these deep VA-based NFs [16] performed better than various multi-modal NF combinations for leadership applications, they did not perform well enough for the classification of low/high extraversion. This shows that a more elaborated method to extract deep VA-based NFs is still needed.

To perform an effective nonverbal behavior analysis, it can be beneficial not to use NFs extracted from all frames (called "holistic approach" in this study) but instead find a way to detect the frames carrying more discriminative information, the so-called "key-frames", and extract NFs only from them. In this work, a deep learning model is proposed to detect key-frames to test this claim. Related to this, in [9], [18], to detect emergent leaders, only the frames having significant motion were used to extract VA-based NFs and other frames were discarded from analysis. This approach is reasonable if some part of the video have little motion such that the corresponding frames are very similar to each other and consequently, the NFs extracted

• *C. Beyan, A. Zunino, M. Shahid and V. Murino are with Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Genoa, Italy. M. Shahid is also with University of Genoa, Electrical, Electronics and Telecommunication Engineering and Naval Architecture Department, Genoa, Italy and V. Murino is also with Department of Computer Science, University of Verona, Verona, Italy and Huawei Technologies in Dublin, Ireland. (e-mail: cigdem.beyan@iit.it; andrea.zunino@iit.it; shahid.muhammad@iit.it; vittorio.murino@iit.it).*

Manuscript received xx xx, 2019; revised xx xx, xx.

from these similar frames are almost the same, i.e. possibly creating redundancy and ultimately misleading the classification task. In [23], a key-frames selection approach which identifies the salient instants of a video was proposed to analyze dyadic human interactions, demonstrating that using features extracted only from key-frames was enough to obtain the desired performance.

In this study, a novel methodology, using deep VA-based NFs extracted only from key-dynamic images, is introduced and applied for the classification of perceived personality traits. We define key-dynamic images as the most informative dynamic images, carrying more discriminative information such that the NFs extracted from them contribute the classification task more than NFs extracted from other dynamic images. The proposed method includes construction of multiple dynamic images [19] to represent short-term VA. Long-term VA is modeled by using a Convolutional Neural Network + Long Short-Term Memory Network (CNN+LSTM) architecture, which allow us to detect the spatio-temporal saliency in each dynamic image. These spatio-temporal saliencies are used to determine the key-dynamic images that VA-based NFs are extracted from. The obtained feature vectors representing the set of key-dynamic images are encoded into a single feature vector using covariance [16]. This resulting feature vector is used for learning and classification with Support Vector Machine.

To the best of our knowledge, this is the first attempt in social and affective computing domains that, a deep learning model is utilized to determine key-dynamic images (also key-frames) by using spatio-temporal saliency. This is also the first study that, dynamic images are combined with CNN+LSTM architecture. Given that dynamic images are a combination of many RGB images, the input video composed of dynamic images becomes much shorter than original video, which results in faster training. The covariance-based feature encoding technique proposed in [16] is also modified such that better classification accuracies are obtained.

Our method, automatic personality traits classification, is tested on a dataset composed of small group interactions during meetings [9] and on a dataset composed of vlogs [24]. The computational study of personality traits is a challenging research topic as demonstrated by nonverbal communication literature since it might involve multiple individuals interacting and personality is simultaneously expressed through many nonverbal channels such as voice, face, body, and so forth [13]. However, in this study we only focus on VA.

The main contribution of this work is presenting an effective deep VA-based NFs extraction mechanism, which results in improved performance as compared to the state-of-the-art VA-based NFs for automatic classification of personality traits. A comprehensive experimental analysis is presented to prove the superiority of the proposed method. As feature representations are automatically learned from the data itself, it is possible to adapt the proposed method to other applications based on nonverbal behavior analysis.

The rest of this paper is organized as follows. The studies using dynamic images, social and affective computing approaches applying key-frames selection algorithms, and related personality traits detection works are reviewed in Section 2. The datasets utilized are described in Section 3. In Section 4, the proposed method is explained in detail. The experimental analysis, and the brief description of the comparative methods are given in Section 5. Following that, the results of comparative study e.g. the proposed method vs. previously published best results of the datasets used (called "baselines" for the rest of this paper), the proposed method vs. different key-frames selection strategies, and the results of ablation study are given in Section 6. Finally, the paper is concluded with discussions and future work in Section 7.

2 Related Work

In this section, studies based on dynamic images [19], works on social and affective computing domains applying key-frames selection algorithms and personality traits detection studies are discussed, and the main differences between our approach and theirs are highlighted.

2.1 Dynamic Images

Dynamic image is a summarization of an input video sequence in terms of the objects in an action (appearances of objects) and summarizes the dynamics of objects' actions (dynamics of motions) by capturing the temporal ordering of the pixels from frame-to-frame. It can be seen as an early fusion of video frames to obtain a compact representation [19].

Dynamic image representation has mainly been used for action, activity, and gesture recognition. For the first time, Bilen et al. [19] proposed dynamic image representation technique using RGB images (raw video frames) as the input and applied this technique for action recognition. That study [19] is discussed in Section 4.1 in more detail. Following that, some modifications of this technique such as dynamic depth image, dynamic depth normal image and dynamic depth motion normal image were constructed from sequence of depth maps in [25]. In that way, the spatial (i.e. posture) and temporal (i.e. motion) information were encoded at different levels for gesture recognition. By using dynamic images, authors [25] showed that it is possible to use pre-trained CNN models without training millions of parameters. Later, Wang et al. [26] proposed using optical flow images as the input to construct dynamic images (instead of using RGB images as the input) for action recognition. In that way, long-range dynamic of actions were captured and better performance as compared to using dynamic images made from RGB images was obtained. In that study [26], input videos were represented by more than one dynamic image. Differently, in [27], RGB and depth sequences were used for gesture segmentation and recognition such that depth images were converted to dynamic images and RGB modality was used with CNN+LSTM to learn long-term spatio-temporal features.

In this study, we capture short spatio-temporal characteristics of a video using multiple dynamic images [19], which shows better performance as compared to using raw video frames (RGB images) or motion energy images (MEIs). Similar to [26], input videos were represented by more than one dynamic image. It is important to highlight that,

for the meeting dataset [9] used, having multiple dynamic images is essential as the videos used are much longer than the action/gesture recognition datasets have. Different from [26] and [27], we use raw images i.e. RGB frames to obtain dynamic images. Using optical flow images as the source could be an alternative, especially to model the longer-range dynamics of an input. However, given that the proposed method includes LSTM, we assume that it is already able to model longer-range dynamics of a motion. Using optical flow images for the construction of dynamic images could be investigated as future work.

2.2 Key-Frames Selection

A major aim of key-frames selection is to find a set of representative frames capturing the salient content of the video [28]. In [28], key-frames selection techniques were divided into four as: *i)* motion analysis-based; computing if there is a change in motion, *ii)* shot boundary-based; selecting first, middle and last frames of each video shot, *iii)* visual content-based (VCB); selecting a key-frame as the frame having a significant change in the color histogram as compared to previously selected key-frame, and *iv)* clustering-based; grouping the frames and, then determining the frame closest to the centroid of each cluster as the key-frame.

The most popular key-frames selection technique applied by social/affective computing studies is VCB (and its modifications). For instance, in [29], one key-frame was selected as having the closest color histogram to the mean RGB histogram of the whole video clip and this frame was used for fine-tuning the AlexNet for emotion recognition. In [30] deep learning and kernel methods were compared for emotion recognition. As inputs to CNN, key-frames detected as in [29] were given. The results were promising for CNN with fine-tuning as compared to CNN from scratch, Support Vector Regression (SVR) and SVR with transfer learning, showing the importance of using key-frames only. In [31], handcrafted video and audio features and image-based features extracted from a pre-trained CNN model for each key-frame selected as in [29], [30], were used for affect recognition. The results showed that using key-frames only were more effective [31]. In [32], the dissimilarity between consecutive frames was computed by comparing the facial representations via appearance-based features. Frames having the largest average distance scores were selected as key-frames. Even though, this method is very simple, manual frame selection performed better than it for visual and multi-modal emotion recognition tasks. Given the popularity of VCB key-frames selection technique, we tested it under different conditions and compared with our method. (see Section 5.2.3 for more details).

In [28], clustering-based key-frame selection showed that summarizing an emotion video by a set of few key-frames in terms of the variability of facial expressions is enough to describe an emotion. This is also efficient and showing that there is no need to process similar several hundred/thousands of frames. In [33] to reduce the processing time and to perform better discrimination for facial expression recognition, when a local positive maximum in the temporal derivatives of the facial actions was detected, corresponding frame was determined as a key-frame, then used for recognition with a temporal classifier. Wang et al. [34] selected one single frame out of hundreds as the key-emotion frame such that wherever a local maximum in audio intensities were detected, it was assumed as a real emotion display and the corresponding frame was defined as a key-frame where visual features were extracted from that frame. The results showed better recognition performance and a significant decrease in computational cost as compared to using all frames for feature extraction. However, choosing one single frame is very restrictive and not always sufficient for all applications. Trabelsi et al. [23] selected key-frames to identify the salient instants of an interaction sequence. Their motivation to apply key-frames selection was to increase the efficiency of their proposed method as the features used were very high dimensional and, thus, computing them for each frame was very costly.

As seen, key-frames selection was applied *i)* to perform better classification as compared to using all frames, which is in line with our purpose and/or *ii)* to increase the efficiency of the proposed method i.e. to decrease the computational cost. Besides, key-frames selection is directly related to video summarization. Video summarization is beyond the scope of this paper, thus, we do not discuss studies related to that, but interested readers can refer to the review paper: [35].

In this study, we define key-frames (more precisely key-dynamic images) as the most discriminative dynamic images such that the NFs extracted from them contributing the classification more than NFs extracted from other dynamic images. We claim that using NFs extracted only from key-dynamic images result in a classification performance better than holistic approach (using NFs extracted from all dynamic images). Similar to [23], our method is based on detecting saliency. However, different from [23], it is not only spatio saliency but also temporal saliency. Additionally, unlike any other methods, we determine key-frames based on a deep learning model.

2.3 Personality Traits Detection

The detection of personality traits has been investigated in many domains such as social media [36], [37], vlogs [24], [38], [39], radio broadcasts [40], using data from smart phones [41], and in small group meetings [14], [15], [16], [42]. Given that, the proposed method is evaluated on datasets composed of small group meetings and vlogs, in Table 1, works only in these contexts are summarized. This summary is in terms of NFs, and the computational methods used and the contributions of each related work are highlighted in the text.

Some studies e.g. [11], [12], [43] analyzed personality traits using self-reported judgments while others e.g. [13], [14], [15], [16], [24], [42], [44], [45], [46] used perception of external observers. The most frequently examined trait is extraversion, which might be due to the fact that its perception is easier such as in meetings and therefore, annotating it is more reliable than the other traits [13].

In [11], it was observed that, the best performing NFs are audio-based i.e. conversational activity and emphasis for extraversion detection. In [12], it was shown that, feature selection on acoustic features by comparing their means

TABLE 1: Related Works: Personality Traits Detection (NA stands for not applicable).

[11]: **Audio-based NFs:** Emphasis (e.g. formant frequency, spectral entropy), Conversational Activity (e.g. energy in frame, length of speaking segments), Influence (e.g. the ratio of overlapping speech segments to the total), Mimicry (e.g. short interjections) **Video-based NFs:** Fidgeting (e.g. head, hands and body movements, motion history images) **Computational Method:** SVM
[12]: **Audio-based NFs:** Conversational Activity, Mimicry, Emphasis, Influence **Video-based NFs:** Head, body and hand fidgeting **Computational Method:** SVM
[43]: **Audio-based NFs:** Speaking activity **Video-based NFs:** Visual Focus of Attention (VFOA) **Computational Method:** SVM
[47]: **Audio-based NFs:** Conversational Activity and Emphasis **Video-based NFs:** VFOA (e.g. attention given or received). **Computational Method:** Naive Bayes, Hidden Markov Models (HMM), and SVM
[44]: **Audio-based NFs:** Speaking Activity, Prosody **Video-based NFs:** NA **Computational Method:** Boosting (weak classifier: one-level decision tree)
[45]: **Audio-based NFs:** Speaking Activity, Prosody **Video-based NFs:** Weighted motion energy image-based (wMEI) features, Facial landmark points **Computational Method:** Random Forest Regression
[48]: **Audio-based NFs:** NA **Video-based NFs:** wMEIs (e.g. mean, median, entropy) **Computational Method:** Ridge Regression (RR), and SVM
[49]: **Audio-based NFs:** Speaking Length **Video-based NFs:** VFOA (e.g. attention given and received) **Computational Method:** SVM
[50]: **Audio-based NFs:** Conversational Activity, Emphasis, Influence, Mimicry **Video-based NFs:** Fidgeting (e.g. head, hands and body movements) **Computational Method:** Bayesian Network
[51]: **Audio-based NFs:** Speaking activity and prosody **Video-based NFs:** VA-based NFs **Computational Method:** SVM and RR
[52]: **Audio-based NFs:** Conversational Activity (e.g. speaking length), Emphasis (e.g. pitch, amplitude) **Video-based NFs:** VFOA (e.g. attention given and received, the same NFs while speaking) **Computational Method:** Generative linear chain HMM, Discriminative Linear-chain Conditional Random Fields, Generative Influence Modeling
[46]: **Audio-based NFs:** Turn-level acoustic NFs (e.g. pitch, intensity, MFCC, etc.) **Video-based NFs:** NA **Computational Method:** Bidirectional Long Short Term Memory with Interlocutor-Modulated Attention Mechanism
[13], [14], [15], [42]: **Audio-based NFs:** Speaking Activity, prosody **Video-based NFs:** Head activity, Body activity, weighted motion energy image (wMEI), and VFOA **Computational Method:** RR and linear SVM
[16]: **Audio-based NFs:** NA **Video-based NFs:** Deep visual activity features **Computational Method:** SVM and Localized Multiple Kernel Learning (LMKL)
NJU-LAMDA [24], BU-NKU [24], ITU-SiMiT [24] : **Audio-based NFs:** Logarithm Mel-filter bank energies in NJU-LAMDA [24] **Video-based NFs:** Face-based nonverbal features using VGG-face and VGG-16 **Computational Method:** Support vector regression BU-NKU [24], Kernel Extreme Learning ITU-SiMiT [24], Softmax NJU-LAMDA [24].

through ANalysis Of VAriance (ANOVA) do not improve the results and it is possible to use thin slices of interactions to detect the extraversion. In [43], [47], [49], different from speaking length, Visual Focus of Attention (VFOA) were detected as an effective indicator of extraversion such that the distribution of attention was an important cue while attention received from peers was insufficient when used alone. Fang et al. [51] demonstrated that speaking interruption information is among the best performing NFs. In [52], mean of attention given, attention received, attention received while not speaking, formant frequency, spectral entropy, and energy in a frame were found as the best performing NFs for extraversion. These findings are in line with the results of previous studies. In [45], it was shown that, NFs different from facial landmark features performs better when used alone, but using NFs all together still performs the best.

Differently, in [44], linguistic features have been used in addition to NFs. The results showed that NFs such as prosody, speech activity overlapping and interruptions out-

perform linguistic features (e.g. words n-gram and dialog acts) for personality trait classification. In [48] personality trait prediction was investigated for cross-domains; from video blogs to small group meetings. Similar to our work, in that work [48], only VA was considered. This was because body activity is one of the key nonverbal cues to signal traits, and VA is a more robust feature when cross-domains are considered, unlike the popular cue; speaking turn taking, which can be highly different in two domains. Lin and Lee [46] utilized audio-based NFs only and applied temporal modeling of vocal behaviors between interlocutors via embedding interaction-based attention mechanism in a BLSTM, which showed better results as compared to several variations of LSTM. That study [46] is different from any other methods reviewed in this section since it uses deep learning. However, the features used are still handcrafted, thus, that method [46] is not data-driven as our method is.

In this study, we investigate classification of personality traits, which are determined based on the judgment of external observers, i.e. perceived personality traits. To evaluate the effectiveness of the proposed method, a meeting dataset [9] and a vlog dataset [24] were used. For the meeting dataset [9], the baseline methods are [13], [14], [15], [16], [42] as they have used the same dataset and also have included VA-based NFs. Except [16], all methods have used handcrafted multi-modal (audio and video-based) NFs. Being based on deep VA-based NFs only, [16] is the most similar work to ours.

On the other hand, for the vlog dataset [24], there is no study investigated VA-based NFs extracted from full visible body; therefore, our study is the first. Video-only baseline methods are BU-NKU [24] and ITU-SiMiT [24]. Both used NFs extracted from face using VGG-face and/or VGG-16, while Support Vector Regression is applied in BU-NKU [24] and Kernel Extreme Learning Machine is applied in ITU-SiMiT [24]. However, the best performing method on average for this dataset is NJU-LAMDA [24], which additionally uses audio-based NFs (logarithm mel-filter bank energies). More information regarding all these baselines with a discussion are given in Section 5.1. Their results are used for comparisons in Section 6.

Unlike any related work, we propose using dynamic images to represent VA in minor intervals and LSTM to model the longer durational VA. CNN+LSTM model is fine-tuned for low/high personality traits classification while allowing us to extract the spatio-temporal saliency of each dynamic image. Dynamic images having high saliency are selected as key-dynamic images and VA-based NFs are extracted only from them for learning and inference.

3 The Datasets

The proposed method was tested on two different dataset, one includes meetings between 3-4 persons and the other composed of vlogs i.e., single person speaking in front of a camera. More details are given as follows.

3.1 ELEA-AV Corpus [9]

ELEA-AV includes 27 meetings having either three or four participants while each meeting lasts 15 minutes on average. The data acquisition set-up was composed of a microphone array and two wide-angle web cameras. The perceived personalities were annotated by external observers using the Ten Item Personality Inventory (TIPI), with a 7-point Likert scale [15]. The annotations are in terms of the average of the scores of annotators for each trait individually. It includes extraversion, agreeableness, conscientiousness, emotional stability and openness.

Following the baselines [13], [14], [15], [16], [42], to decide the labels as low or high extraversion, agreeableness, conscientiousness, emotional stability or openness, for each participant in a given meeting, the median of the corresponding trait's scores are used. The participant having score smaller/greater than the median score of the corresponding personality trait is labeled as low/high.

3.2 ChaLearn First Impression Dataset [24]

This dataset contains 10000 clips extracted from 3000 different YouTube vlogs, i.e. videos of people facing and speaking in English to a camera. Average duration of videos is 15 seconds. The people have different gender, age, nationality and ethnicity, which make the task of inferring apparent personality traits more difficult. Ground-truth estimation was performed in terms of the perception of human subjects visioning the videos labeled with personality traits variables. As personality traits, the Five Factor Model (the Big Five), that models human personality along five dimensions: extraversion, agreeableness, conscientiousness, neuroticism and openness were used. Thus, each clip has ground truth labels for these five traits represented with a value within the range [0, 1]. They were turned to binary classes as high or low, by thresholding them at 0.5 [24].

This dataset is a part of workshop series such that it has been introduced as a challenge. In the first phase of the challenge, all teams (competitors, i.e. baseline methods) used 6000 videos for training and 2000 videos for test phase. In the second phase of the challenge, another set of 2000 videos were used for test phase. However, in the second phase of the challenge, not all the baselines used the same videos for training (e.g. some used validation set for hyper parameter selection, others combined validation set with training set such that they trained their model with a larger set, or applied additional cross validation techniques). In this study, we use the full training set and the full test set of first phase (so-called validation set) to guarantee that the training and test videos used and the way cross validation applied are the same with baselines. The reported results of first phase in [24] also includes classification of the personality traits (i.e. not only regression), which allows us to make fairer comparisons.

4 Proposed Method

The proposed method can be summarized as follows. Given a training video, first, dynamic images are obtained as described in Section 4.1. Then, these images are used to fine-tune a CNN+LSTM model as explained in Section 4.2. This trained CNN+LSTM model provides spatio-temporal saliency of each dynamic image while the images having higher spatio-temporal saliency are selected as key-dynamic

images. Later, CNN fully connected layer 6 (fc6) features of key-dynamic images (VA-based NFs) are combined using improved covariance-based feature encoding technique (I-CFE, Section 4.3) such that a single feature vector (called final NFs) is obtained. Final NFs are used to learn an SVM model to perform personality traits classification.

For a new video (i.e. test video), dynamic images are obtained. Then, key-dynamic images are automatically determined using the trained CNN+LSTM model. The VA-based NFs of key-dynamic images are encoded by using I-CFE, and the resulting final NFs are classified by SVM trained. An illustration of the proposed method is given in Figure 1.

4.1 Construction of Multiple Dynamic Images

Dynamic image is a compact representation of a video, which summarizes the appearance and dynamics of it. Construction of a dynamic image contains rank pooling that encodes the temporal evolution of the frames in a video and potentially enables the use of any CNN model with fine-tuning. Given that dynamic image representation highlights the object in an action, it is a direct competitor of Motion History and/or Motion Energy Images [17], which were utilized in [9], [11], [13], [14], [15], [16], [18] to extract VA-based NFs. The details of its algorithm can be found in [19].

A dataset used might contain long videos i.e. more than 20000 frames (e.g. ELEA-AV [9]) and using deeper networks can be reasonable to model such long sequences, which consequently involve huge number of parameters and might create problems such as not fitting in the GPU memory, and high computational complexity. The actions needed to handle these problems are such as clipping the temporal receptive fields of the videos to short durations, which prohibit CNNs and LSTMs from learning long-term temporal data [27]. In this study, thanks to using multiple dynamic images such problems were encountered less.

For each consecutive N frames in a video (N is taken as 100 for ELEA-AV dataset [9] and 20 for ChaLearn dataset [24]), we obtain one dynamic image with no overlapping (such as 27597 number of frames were represented by 276 dynamic images). The input video for feature learning composed of dynamic images becomes much shorter than original video and this results in faster training. Using shorter sequences is even more advantageous when LSTM is used as in this way it is possible to prevent problems, e.g. vanishing gradients occurring when the input sequences are very long.

We also observed that dynamic images constructed from RGB domain were good at capturing the motions belong to the person in a video. The results of this step were well enough so that we did not need to try other domains such as optical flow images as the source of the construction of dynamic images. Once dynamic images are extracted, they are used to fine-tune a deep model for low/high personality trait classification as described in Section 4.2.

4.2 Nonverbal Feature Extraction and key-Dynamic Images Selection

Given a video sequence composed of dynamic images, to detect the key-dynamic images and extract VA-based NFs from them, we adapt a very recent work [21]. That method [21] was originally proposed for action recognition and video captioning. It devises a probabilistic formulation simultaneously grounding evidence in space and time for a video input. It also visualizes the spatio-temporal cues that contribute to a deep recurrent neural network model trained.

Motivated from that work [21], a CNN+LSTM architecture is trained for personality traits classification (e.g.; low/high extraversion), to exploit the activated neurons in a top-down backpropagation framework. During test phase, a video sequence composed of dynamic images are given as the input to this trained architecture to compute a forward pass, activating excitatory paths, specific to different classes. During the backward pass, we isolate the excitatory connections by setting a one-hot vector according to the ground truth class on the top of the CNN+LSTM, and backpropagate this one-hot vector through the network (see Figure 1). The probabilistic top-down backpropagation framework results in a series of saliency maps, each corresponding to one dynamic image. Spatio-temporal saliency maps are used to localize where (in spatial dimension) and when (in time) the discriminative cues exist for low/high personality trait classification as follows. The saliency maps are summed up at the pixel level to obtain compact information for each dynamic image. Following the insights reported in [21] (namely the pointing game in time), in this study, we claim that the time position of the peak in the sum of the saliency maps can be used to locate the images that the model utilize the most in order to distinguish classes (i.e. the dynamic images contributing the classification task more than other dynamic images, namely, key-dynamic images). Grounding on this claim, we sort the sums of the saliency maps (from the highest to the lowest values) in time, which results in an image ranking, showing the contribution of each image to the classification task from the most to the least. Using this ranking, the images having high saliency are selected as key-dynamic images.

We fine-tune the AlexNet architecture (pre-trained on ImageNet) as the CNN part of our model, which has two dropout layers, each after the fully connected layers. fc6 features extracted for each dynamic image are recursively given as an input to a single layer LSTM (256 hidden units), which is jointly trained with the AlexNet. A dropout layer is followed the LSTM layer. All dropout layers had 0.5 dropout rate. We split the training videos in 32 consecutive dynamic images long clips for ELEA-AV [9] and 15 consecutive dynamic images long clips for ChaLearn [24] and train the AlexNet+LSTM for 30000 iterations, using stochastic gradient descent optimizer. The learning rate was set to 10^{-3} for the AlexNet layers while to 10^{-2} for the LSTM layer.

The number of key-dynamic images is set to 50 for ELEA-AV [9] (approximately 20% of the total number of dynamic images constituting a video) and 15 for ChaLearn [24] (approximately 65% of the total number of dynamic images constituting a video). We also tested using the same amount of dynamic images selected by the comparative method given in Section 5.2.3. fc6 features are extracted for each key-dynamic images, which are later combined using improved covariance-based feature encoding (I-CFE) before

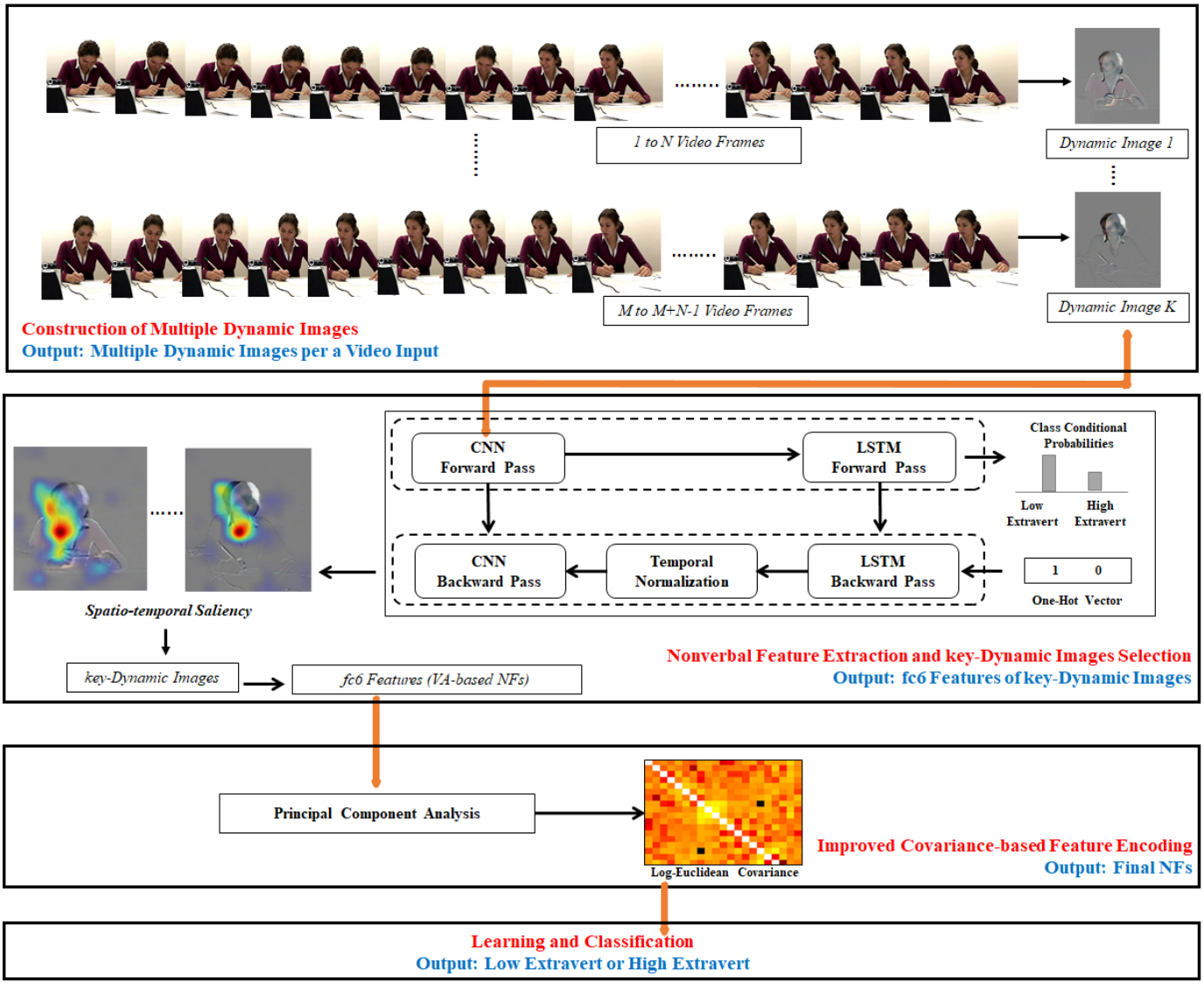

Fig. 1: An illustration of the proposed method. For the sake of simplicity, extraversion task is given as an example. The name of each component of the proposed method is written in red, and the output of each component is written in blue. Given a training video, first multiple dynamic images are obtained. Then, these images are used to fine-tune a CNN+LSTM architecture, which results in spatio-temporal saliency for each dynamic image. Dynamic images having higher spatio-temporal saliency are chosen as key-dynamic images and their fc6 features are combined using covariance-based feature encoding method, after applying PCA. The resulting feature vector is used to train an SVM classifier. During test phase, dynamic images of test video are obtained, the CNN+LSTM trained is used to find the key-dynamic images. PCA is applied to fc6 features of key-dynamic images, which are encoded by using I-CFE afterwards. The resulting final feature vector is classified using the trained SVM model as low extravert or high extravert.

applying learning and classification.

4.3 Improved Covariance-based Feature Encoding (I-CFE)

Following the work [16], we first apply Principal Component Analysis (PCA) to reduce the dimensionality of the fc6 feature vectors and to remove the correlations between features. For the experimental analysis with ELEA-AV dataset [9], unlike [16], which defined the number of PCA used as the smallest number of components that represents 90% of the sum of all eigenvalues, we fixed it to 10 empirically. This is because using the same rule with [16] resulted in still high dimensionality (e.g. 360 dimensions), which might be due to the fact that fc6 features obtained were very sparse that could be a result of using dynamic images. For ChaLearn dataset [24], we found the optimum number of components as in [16]. Later, the covariance matrix is calculated as follows.

$$C_p = \frac{1}{n-1}\sum_{k=1}^{n}(f_p(k)-\mu_p)(f_p(k)-\mu_p)^T \qquad (1)$$

where p is a video, μ_p is the mean of the PCA applied, f_p is the fc6 feature vectors, and n is the total number of the feature vectors (i.e. the number of dynamic images). The diagonal entries of covariance matrix are composed of the

variance of each feature and the non-diagonal entries are composed of the correlations.

The matrix logarithm operation is used to map covariance matrices from the Riemannian manifold to the Euclidean space. To do that, singular value decomposition, which decomposes the covariance matrix, is applied as follows.

$$C_p = U\Sigma U^T \tag{2}$$

where U is a $d \times d$ orthonormal matrix, and Σ is the square diagonal matrix with nonnegative real numbers while eigenvalues are on the diagonal. Then, the new representation of the covariance matrix is written as:

$$\epsilon = 10^e \tag{3}$$

$$Sigma_{new} = \Sigma + \epsilon I \tag{4}$$

$$C_p^{(log)} = \log(C_p) = USigma_{new}{}'U^T \tag{5}$$

where I represents the identity matrix with the same size as C_p. e was taken as= {0, -1, -2, -3, -4, -5, -6, -7, 1, 2, -0.5, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7} Σ' is the $d \times d$ square matrix with logarithmic values of the eigenvalues on the diagonal. The calculation of $Sigma_{new}$ is the part added in this study.

The final NFs vector obtained as the result of this encoding is the all entries on and above (below) the diagonal of the covariance matrix.

4.4 Learning and Classification

The final NFs vector obtained by applying I-CFE is used for learning and inference. Given that the most popular classifier among baselines is Support Vector Machine (SVM) (see Section 5.1), we apply SVM as well. In this scope, the radial basis kernel function (RBF) and linear kernel were used. As kernel parameters C was taken as 2^i, $i = -1, 1, 3...31$ and RBF γ was used as 2^j, $j = -11, -9, -7...11$ to be in line with [16].

5 Experimental Analysis

For ELEA-AV dataset [9], SVM was applied with leave-one-out cross validation, and accuracy (ACC) was used to evaluate the performance of the proposed method. ChaLearn dataset [24] supplies separate videos to be used in the training and test phases. Thus, we used them, accordingly. For ChaLearn dataset [24], area under curve (AUC) was used as the evaluation metric. It is important to highlight that the evaluation metrics used, the data partition protocol and the method to divide the dataset into two classes for each personality traits are in line with baseline methods of each dataset.

5.1 Baselines

The state-of-the-art methods for the datasets used in this study are briefly described as follows and the corresponding performance of each method is given in Tables 2 and 3.

The baselines of ELEA-AV dataset [9] are as follows.

- Aran et al. [13] uses various audio and visual NFs (in total 73) based on: speaking activity, prosody, head activity, body activity, weighted motion energy image (wMEI), and VFOA. Linear SVM and Random Forest (RF) are applied to detect the low/high personality traits.
- Using exactly the same NFs with [13], Okada et al. [14] presents finding co-occurring segments of NFs in the data, which are determined based on graph clustering and in that way time-series signals are converted into binary events. Linear SVM is used for classification and on average, their approach results in better performance than [13].
- In [15], the same NFs with all the works above are utilized. Maximum Relevance Minimum Redundancy (MRMR) feature selection method and Principal Component Analysis (PCA) are applied with the classifiers: SVM, and RF. Similar results were obtained using each classifier while applying MRMR performs better than [13], [14] when 50 NFs out of 73 NFs used. In the same study [15], multi-task learning by using emergent leadership as a trait and transfer learning by utilizing other extraversion datasets, has been also tested and showed improved performances as compared to [13], [14].
  All these studies use handcrafted NFs and represent VA in terms of head activity detected using optical flow, body activity detected using image differencing and head/body activity based on weighted motion energy image (wMEI). They all showed that, generally, the best performing NFs are VA-based, although using all NFs together resulted in the better accuracy.
- Recently, the modified version of [14] was published as [42]. The early fusion applied in [42] tends to create imbalance in the number of NFs between modalities such that the number of co-occurrence NFs based on categorical events are larger than the number of NFs based on binary events. To avoid this, in [42], late fusion is applied and overall better results as compared to [14] were obtained.
- Different from all these studies discussed above, in [16] a Hybrid CNN model [20], [22], which takes optical flow images as inputs, is used to extract deep VA-based NFs for each frame. To represent a video, NFs per frame are combined using statistical feature encoding or covariance-based feature encoding into a single feature vector. As the classifier, SVM and Localized Multiple Kernel Learning (LMKL) are used. Deep VA-based NFs performs better than all handcrafted VA-based NFs given above when LMKL is used, but this does not perform as well as the best method i.e. [15] for low/high extraversion classification. As mentioned before, in our study, we have integrated covariance-based feature encoding applied in [16] to our methodology with a small modification that results in better performance not only for the proposed method but also for method introduced in [16] (for results, see Section 6.5).

As the baselines of ChaLearn dataset [24], we have included the best performing approach on average, which is a multi-modal approach, and all video-based approaches. The details of them are as follows. The descriptions of other baselines and corresponding performances can be found in [24].

- NJU-LAMDA [24] uses audio and face-based NFs. Audio-based NFs are logarithm Mel-filter bank energies. The face-based NFs are extracted using VGG-face. Late fusion of NFs is applied. Among all baselines, that method performs the best on average for the classification of

personality traits.

- BU-NKU [24] uses face-based NFs only, which are extracted using VGG-face pre-trained on Fer2013 dataset and applies face alignment and separation of scene and face, before extracting features. Kernel Extreme Learning Machine is used for the classification task.
- ITU-SiMiT [24] uses VGG-face and VGG-16 for the extraction of face-based NFs after applying face detection and alignment. SVM is used for the classification.

It is important to recall that the baselines of ChaLearn dataset [24] are customized specifically for that dataset. For instance, they all focus on face and some of them also separate face from background. This requires applying pre-processing steps e.g., face detection, frontalization, and alignment. Whereas, our method is more general e.g., not requiring a specific body part detection and does not perform any pre-processing step. We also do not separate background from foreground assuming that dynamic image creation can handle it automatically. On the other hand, ChaLearn dataset [24] contains very short videos (15 seconds) as compared to ELEA-AV dataset [9]. We selected more key-dynamic images than we did for ELEA-AV dataset [9], as the same percentage of key-dynamic images was not enough to obtain competitive results.

5.2 Comparative Methods

In this section, the implementation details of the methods used for ablation study and experimental comparisons are given. They were all applied for low/high extraversion classification using ELEA-AV dataset [9].

5.2.1 Motion Energy Image (MEI)

Motion energy image is an alternative way to represent visual activity. It was used in baselines: [13], [14], [15], [42] as well.

We integrated MEI into the proposed pipeline such that instead of using dynamic images, MEIs were used. The algorithm of MEI can be found in [17]. We used pixel threshold as 40 (set by visual observation) and for every 100 consecutive frames, one MEI was obtained, which is in line with dynamic images creation (experiments with ELEA-AV dataset [9]). The first frame of every 100 frames was taken as initial image as required by MEI algorithm. The resulting MEIs are binary, but AlexNet requires three-channel images. Therefore, we replicated the binary image for each color channel, after converting them to unsigned 8-bit integer and then multiplying by 255.

5.2.2 Fine-tuning AlexNet

A simpler approach as compared to fine-tuning AlexNet+LSTM can be fine-tuning AlexNet when multiple dynamic images are the inputs. Except the task (i.e. personality traits classification), this is similar to [19].

There are number of ways to fine-tune a deep model. One common way is updating the fully connected layers only. We followed this approach such that the last fully connected layers i.e. fc7 and fc8 were updated while the earlier layers' weights were kept as they were. The model was trained using Adam optimizer with fixed learning rate $1e^{-5}$ and 0.50 dropout rate (experimentally set) was used for all fully connected layers. For each cross-validation fold, training were continued until 40 epochs.

5.2.3 Visual Content-based key-Frames Selection

For M consecutive frames, one frame (or dynamic image) is selected as the key-frame (or key-dynamic image). In detail, the mean of color histograms of M consecutive frames is calculated and the distance between this histogram and the color histogram of each frame is calculated using Manhattan distance. The frame having the closest histogram to the mean histogram is selected as the key-frame (key-dynamic image). M was taken as 32, which is equal to the number of LSTM cells used to test the proposed method (Section 4.2). This method was applied with dynamic images and RGB images (raw video frames) individually, and referred as KeySOA in the following sections. This method was tested for low/high extraversion classification using ELEA-AV dataset [9].

5.2.4 Uniformly Random Frames Selection

Uniformly random frames (or dynamic images) were selected when dynamic images or RGB images were the input. The number of frames to be selected was set to 50 or to the same amount of frames selected by KeySOA. Random frames selection was tested for low/high extraversion classification using ELEA-AV dataset [9].

5.2.5 Applying Softmax as the Classifier

Instead of extracting fc6 features, combining them with I-CFE (Section 4.3) and then applying SVM, it is also possible to perform the classification using Softmax function, which can make the proposed method more end-to-end. This was applied as follows.

Once the key-dynamic images (or key-frames) are obtained, it is possible to combine Softmax score of them to obtain the final class prediction as low or high extravert. This was applied (as performed in [16]) *i)* by averaging the class probabilities, which are obtained from softmax transformation for each key-dynamic image (or key-frame) of a video and then deciding the class of a video as the class having the highest probability, or *ii)* by applying majority voting to the predicted labels of all key-dynamic images (or frames). Whichever method resulted in better performance than other is given in Section 6 as Softmax.

6 Results

This part includes comparisons between the proposed method and the baselines (Section 6.1) of two different datasets used. Furthermore, the results of the ablation study (Section 6.2), comparisons between the proposed method and the other key-frames selection strategies (Section 6.3), visualization of the spatial saliency (Section 6.4) and the results with/without the modification proposed for the calculation of covariance-based feature encoding (Section 6.5) are given. These latter experimental analyses were applied using ELEA-AV [9] for low/high extraversion classification.

6.1 Comparisons with Baselines

The best results of the baselines and the proposed method using ELEA-AV [9] and ChaLearn [24] datasets, are given in Tables 2 and 3, respectively. For the proposed method, the results given in Table 2 corresponds to using 50 key-dynamic images and in Table 3 corresponds to 15 key-dynamic images. Interested readers can refer to supplementary material for the percentage of true positives, true negatives, false positives and false negatives of the proposed method for each trait.

As can be seen (Table 2), on average, the proposed method performs (72%) much better than [13] (57%) [14] (60%) and [42] (63%), although these studies use various multi-modal NFs together.

For extraversion classification, the proposed method performs (77%) better than [15] when only Head/BodyAct-based NFs (67%) are used and much better than [15] when only wMEI-based NFs (66%) are used. Additionally, the proposed method performs as well as the best performing baseline (i.e. [15]) (77%). This result is important as that baseline [15] utilizes 73 different multi-modal features and applies feature selection, while our approach is VA-based only and we do not apply feature selection. The proposed method also performs better than the deep VA-based NFs classified with Localized Multiple Kernel Learning (LMKL) in [16] (72%). LMKL often performs much better than SVM as being based on multiple kernels and be able to automatically assign lower weights to the features not contributing much to the classification task (see [10] for more information). Therefore, if the proposed method is combined with LMKL instead of SVM, it is highly possible that even better classification results would be obtained.

For agreeableness and conscientiousness traits classification, the proposed method performs 13% better than the best performing baseline of each traits i.e. [14] and [15], respectively. These baselines [14], [15] also use multi-modal NFs and additionally [14] requires co-occurrence calculation, thus, the performance of the proposed method is a very important achievement. For emotional stability classification, the proposed method performs 6% better than the best performing baseline (i.e. [15]), which is based on multi-modal NFs. Finally, for openness to experience classification, the proposed method performs 3% better than the best performing baseline (i.e. [42]), which is based on late fusion of multi-modal co-occurring NFs.

Better average performance of the proposed method is definitely a success, but having low standard deviation (5%) of all personality traits while still performing better on average, is also a significant aspect of the proposed method. Lower STD value of the proposed method shows that its performance is generalizable. Its performance (STD=0.05) is also more consistent as compared to the best performing baseline ([42], STD=0.07). In detail, the proposed method performs the best for the classification of agreeableness (79%) and extraversion traits (77%) while its worst performance is for the classification of emotional stability trait (67%).

As can be seen (Table 3), on average, the proposed method performs (79%) much better than ITU-SIMIT [24] (49%), which uses face-based NFs. Its average performance (79%) is as good as BU-NKU [24] (80%) that is based on face features and worse than NJU-LAMDA [24] (82%), which utilizes audio with face features. The proposed method performs better than all baselines for agreeableness trait (3% better than the best performing baseline), while its worst performance is for openness trait (5% worse than the best performing baseline). Similar to the results obtained with ELEA-AV dataset [9], the performance of the proposed method is more consistent (STD=0.01) as compared to all baselines.

6.2 Ablation Study

In this part, given the proposed methodology, we evaluate *i)* whether using dynamic images instead of using RGB images (raw video frames) or MEIs performs better or not, *ii)* if using LSTM supplies an improvement or not, *iii)* whether selecting key-dynamic images has a positive contribution or not and *iv)* if using I-CFE with SVM instead of using Softmax as the classifier has a benefit or not. All these analyses were performed using ELEA-AV dataset [9] for low/high extraversion classification.

We fine-tuned AlexNet for low/high extraversion classification as described in Section 5.2.2, when dynamic images were the inputs, then extracted fc6 features from all dynamic images, combined them using I-CFE (as described in Section 4.3) and finally applied SVM (as described in Section 4.4). Corresponding result is shown as DynImag-AlexNet-SVM in Table 4. This experiment allows us to understand the contribution of LSTM and intrinsically key-dynamic images selection.

We applied holistic approach such that first, AlexNet+LSTM was fine-tuned (as described in Section 4.2), then fc6 features of all dynamic images were combined using I-CFE (meaning that spatio-temporal saliency detection and key-dynamic images selection were not applied) and lastly, SVM was performed. This was applied using MEIs, RGB and dynamic images as the inputs and corresponding results are shown as MEI-AlexNet+LSTM-SVM, RGB-AlexNet+LSTM-SVM and DynImag-AlexNet+LSTM-SVM, respectively in Table 4. The same pipeline, by removing I-CFE with SVM but applying Softmax as the classifier (as described in Section 5.2.5), were also tested and corresponding results are shown as MEI-AlexNet+LSTM-Softmax, RGB-AlexNet+LSTM-Softmax and DynImag-AlexNet+LSTM-Softmax in Table 4. These experiments were carried out to understand the contributions of key-dynamic images selection vs. holistic approach, I-CFE method with SVM vs. Softmax and dynamic images vs. RGB images or MEIs.

Additionally, the best result of the proposed method (obtained when 50 key-dynamic images were used), shown as Proposed Method, and applying all components of the proposed method except I-CFE with SVM but instead Softmax, shown as DynImag-AlexNet+LSTM-Key-Softmax, are given in Table 4. With this comparison, we were able to observe the contribution of I-CFE method with SVM vs. Softmax.

The best result of applying all components of the proposed method except multiple dynamic images construction (i.e. using RGB raw images or MEIs as the

TABLE 2: Comparisons with the baselines of ELEA-AV dataset [9] for personality traits classification. The evaluation metric is accuracy. NA, Ext, Agree, Cons, EmoSt, Open, AVG, and STD refer to not available, extraversion, agreeableness, conscientiousness, emotional stability, openness to experience traits, the average and the standard deviations of the performances, respectively.

Method	Details	Ext	Agree	Cons	EmoSt	Open	AVG	STD
[13]	Multi-modal NFs, SVM	0.63	0.53	0.53	0.54	0.62	0.57	0.05
[13]	Multi-modal NFs, RF	0.62	0.49	0.46	0.51	0.49	0.51	0.06
[14]	Multi-modal NFs, early fusion of co-occurrence fea., SVM	0.68	0.66	0.51	0.57	0.57	0.60	0.07
[15]	Head/BodyAct-based NFs, SVM	0.65	NA	NA	NA	NA	NA	NA
[15]	Head/BodyAct-based NFs, RF	0.67	NA	NA	NA	NA	NA	NA
[15]	wMEI-based NFs, SVM	0.61	NA	NA	NA	NA	NA	NA
[15]	wMEI-based NFs, RF	0.66	NA	NA	NA	NA	NA	NA
[15]	Multi-modal NFs, SVM	0.75	0.59	0.57	0.61	0.54	0.61	0.08
[15]	Multi-modal NFs, RF	0.74	0.52	0.51	0.50	0.53	0.56	0.1
[15]	Multi-modal NFs, SVM, w/feature selection, SVM	**0.77**	NA	NA	NA	NA	NA	NA
[15]	Multi-modal NFs, SVM, w/feature selection, RF	**0.77**	NA	NA	NA	NA	NA	NA
[42]	Multi-modal NFs, late fusion of co-occurrence fea., SVM	0.72	0.65	0.55	0.58	0.66	0.63	0.07
[42]	Multi-modal NFs, late fusion of co-occurrence fea., RF	0.63	0.64	0.54	0.59	0.56	0.59	0.04
[16]	Deep VA-only, Softmax	0.56	NA	NA	NA	NA	NA	NA
[16]	Deep VA-only, SVM	0.64	NA	NA	NA	NA	NA	NA
[16]	Deep VA-only, LMKL	0.72	NA	NA	NA	NA	NA	NA
Proposed Method	Deep VA-only, 50 key-dynamic images	**0.77**	**0.79**	**0.70**	**0.67**	**0.69**	**0.72**	0.05

TABLE 3: Comparisons with the baselines of ChaLearn dataset [24] for personality traits classification. The evaluation metric is Area Under Curve. AVG, STD, Ext, Agree, Cons, Neu, Open refer to the average and the standard deviation of the performances, extraversion, agreeableness, conscientiousness, neuroticism and openness traits, respectively.

Method	Details	Ext	Agree	Cons	Neu	Open	AVG	STD
Random Guess	-	0.50	0.51	0.52	0.50	0.52	0.51	0.01
NJU-LAMDA [24]	Multi-modal NFs	0.83	0.76	0.85	**0.81**	**0.82**	**0.82**	0.03
BU-NKU [24]	Face-based NFs	**0.84**	0.74	**0.86**	0.79	0.80	0.80	0.05
ITU-SiMiT [24]	Face-based NFs	0.44	0.47	0.48	0.49	0.47	0.49	0.02
Proposed Method	VA-based NFs, 15 key-dynamic images	0.81	**0.79**	0.80	0.79	0.77	0.79	0.01

TABLE 4: Ablation study using ELEA-AV dataset [9] for low/high extraversion classification. ACC stands for accuracy. Proposed method refers to DynImag-AlexNet+LSTM-Key-SVM.

	ACC
DynImag-AlexNet-Softmax	0.53
DynImag-AlexNet-SVM	0.70
RGB-AlexNet+LSTM-Softmax	0.46
RGB-AlexNet+LSTM-SVM	0.70
MEI-AlexNet+LSTM-Softmax	0.56
MEI-AlexNet+LSTM-SVM	0.63
DynImag-AlexNet+LSTM-Softmax	0.53
DynImag-AlexNet+LSTM-SVM	0.72
RGB-AlexNet+LSTM-Key-Softmax	0.48
RGB-AlexNet+LSTM-Key-SVM	0.72
MEI-AlexNet+LSTM-Key-Softmax	0.56
MEI-AlexNet+LSTM-Key-SVM	0.69
DynImag-AlexNet+LSTM-Key-Softmax	0.54
Proposed Method	**0.77**

input), shown as RGB-AlexNet+LSTM-Key-SVM, MEI-AlexNet+LSTM-Key-SVM, respectively, and additionally not applying I-CFE with SVM but applying Softmax as the classifier, shown as RGB-AlexNet+LSTM-Key-Softmax and MEI-AlexNet+LSTM-Key-Softmax are given in the same table. In that way, we evaluated the contribution of I-CFE method with SVM vs. Softmax and using dynamic images vs. RGB images and MEIs. The corresponding best results were obtained when 50 RGB frames were selected for all.

As seen in Table 4, dynamic images as the input performs better than RGB images and MEI, for all methods no matter Softmax or SVM is applied. Additionally, even using AlexNet only (ACC=0.70 with SVM) results in better than [16] (ACC=0.64 with SVM in Table 2). The results shows that using I-CFE with SVM is better than using Softmax for all cases, and combining LSTM with AlexNet improves classification performance when SVM is used. Additionally, key-dynamic images selection improves the results (5%) such that the proposed method performs the best of all.

6.3 Comparisons with Different Key-Frames Selection Strategies

The proposed method was compared with the most popular key-frames selection algorithm in social/affective computing (shown as KeySOA in Table 5), as described in Section 5.2.3 by combining it with:

- fine-tuning AlexNet for low/high extraversion classification when dynamic images as the inputs (shown as DynImag-AlexNet-KeySOA in Table 5) and with the classifier either SVM (with I-CFE) or Softmax (as described in Section 5.2.5),
- fine-tuning AlexNet+LSTM for low/high extraversion when RGB frames or dynamic images as the inputs using SVM (with I-CFE) or Softmax as the classifier (shown as AlexNet+LSTM-KeySOA in Table 5).

Additionally, the proposed method was compared with uniformly random dynamic images/frames (shown as uniRndm in Table 5) selection while the number of dynamic images/frames used were as much as KeySOA (shown as 'as KeySOA' in Table 5) or were arbitrary set to 50 (shown as 50 DynImag/frames in Table 5). Uniformly random selection was applied for various combinations i.e. with fine-tuning AlexNet or AlexNet+LSTM, using RGB images or

TABLE 5: Comparisons with different key-frames selection strategies using ELEA-AV dataset [9] for low/high extraversion classification. ACC stands for accuracy. Proposed method refers to DynImag-AlexNet+LSTM-Key-SVM.

	ACC	Details
DynImag-AlexNet-KeySOA-Softmax	0.53	as KeySOA
DynImag-AlexNet-KeySOA-SVM	0.64	as KeySOA
RGB-AlexNet+LSTM-KeySOA-Softmax	0.46	as KeySOA
RGB-AlexNet+LSTM-KeySOA-SVM	0.67	as KeySOA
DynImag-AlexNet+LSTM-KeySOA-Softmax	0.53	as KeySOA
DynImag-AlexNet+LSTM-KeySOA-SVM	0.67	as KeySOA
RGB-AlexNet+LSTM-uniRndm-Softmax	0.46	50 frames
RGB-AlexNet+LSTM-uniRndm-SVM	0.66	50 frames
RGB-AlexNet+LSTM-uniRndm-Softmax	0.46	as KeySOA
RGB-AlexNet+LSTM-uniRndm-SVM	0.68	as KeySOA
DynImag-AlexNet+LSTM-uniRndm-Softmax	0.53	50 DynImag
DynImag-AlexNet+LSTM-uniRndm-SVM	0.68	50 DynImag
DynImag-AlexNet+LSTM-uniRndm-Softmax	0.53	as KeySOA
DynImag-AlexNet+LSTM-uniRndm-SVM	0.70	as KeySOA
RGB-AlexNet+LSTM-Key-Softmax	0.48	50 frames
RGB-AlexNet+LSTM-Key-SVM	0.72	50 frames
RGB-AlexNet+LSTM-Key-Softmax	0.46	as KeySOA
RGB-AlexNet+LSTM-Key-SVM	0.67	as KeySOA
DynImag-AlexNet+LSTM-Key-Softmax	0.54	50 DynImag
Proposed Method	**0.77**	50 DynImag
DynImag-AlexNet+LSTM-Key-Softmax	0.54	as KeySOA
Proposed Method	0.74	as KeySOA

dynamic images, and applying I-CFE with SVM or Softmax. Similarly, the proposed method was tested when the same amounts of frames with KeySOA were selected and when the number of frames selected was set to 50. All these analyses were performed using ELEA-AV dataset [9], for low/high extraversion classification.

In general, for the proposed method and the methods using the same key-dynamic images (or frames) selection approach of the proposed method (i.e. all experiments in third part of Table 5), performs better when 50 dynamic images/frames were used. Whereas, for any combination with uniformly random selection, using dynamic images/frames as much as KeySOA selected results in better performance than using 50 dynamic images/frames. The results of KeySOA were similar to uniformly random selected frames, while random selection even performs better than KeySOA for DynImag-AlexNet+LSTM-SVM. Shortly, results proved that the proposed key-dynamic images selection method i.e. using spatio-temporal saliency is better (ACC = 0.77 and 0.74) than KeySOA and uniformly random selection, while its adaptation when the input images are RGB (i.e. selecting key-frames as proposed instead of key-dynamic images) results in the second best classification performance (ACC = 0.72).

6.4 Visualization of Spatial Saliency for the key-Dynamic Images

For two example videos of ELEA-AV dataset [9] having different ground-truth labels (low or high extraversion), the first 10 key-dynamic images selected by the model (from the most discriminative to the least) during test phase, are given in Figure 2. Each spatial saliency determined by the model is overlaid with the corresponding key-dynamic image. This supplies a visual insight on where (in pixel level) the model had more attention and found discriminative spatial cues to predict the correct class. However, it is important to recall that our model has a spatio-temporal nature. Therefore, this spatial saliency visualization is useful only to understand the learned model at the pixel level.

As can be seen in Figure 2, the trained model (Section 4.2) detects regions having VA such as hands, head, shoulder, etc. The model also eliminates the static background, as it never finds the pixels belong to the static background as discriminative. This is important because this proves that motion is being used as the discriminative factor.

It is hard to distinguish two classes; low or high extraversion, by using spatial information only. For instance, one cannot observe that for low extraversion, the model focuses more on head, while for high extraversion; it focuses more on hands, or etc. This is the evidence that our approach is able to differentiate the two classes by looking at the distribution of the intensity of the saliency in time, and not simply looking at the body part content (e.g. hands or head) i.e. not simply learning appearance of the body parts nor the location of the body parts. In other words, the spatio-temporal characteristic of the model is important and using spatial information only (i.e. using AlexNet only) is not enough, as also proved in Section 6.2 with a quantitative analysis.

TABLE 6: Covariance-based feature encoding calculation using ELEA-AV dataset [9] for low/high extraversion classification. ACC stands for accuracy.

	ACC	Details
[16]-SVM	0.64	e = 0, i.e. as presented in [16]
[16]-SVM	0.66	e = -1 with RBF, e=-6 with linear
[16]-LMKL	0.67	e = 0, i.e. as presented in [16]
[16]-LMKL	0.67	e = -4
Proposed Method	0.74	e = 0, i.e. as presented in [16]
Proposed Method	**0.77**	e = 2

6.5 Modified vs. Original Covariance-based Feature Encoding

In Table 6, the best results which were obtained by using covariance-based feature encoding as proposed in [16] and the best results which were obtained by using our modified version, namely, improved covariance-based feature encoding (I-CFE), are given for the proposed method and for [16] when SVM and LMKL were used. It is worth to note that the best result (ACC=0.72 as given in Tables 2 and 4) of [16] were obtained when another encoding technique was used with LMKL, hence, it is different from the result given here. These analyses were performed using ELEA-AV dataset [9] for low/high extraversion classification.

Results (Table 6) show that most of the time, the proposed modification for the calculation of Eq. 3 (i.e. using the e value) resulted in better classification performance. Herein, as mention in Section 4.3, we tested empirical values for e but, it might be better to learn this value during training which will be investigated as future work.

7 Discussions and Future Work

We have introduced a new methodology to extract visual activity (VA)-based nonverbal features (NFs). The short-term VA of a person is represented by dynamic images [19]. The longer-term VA is modeled by fine-tuning

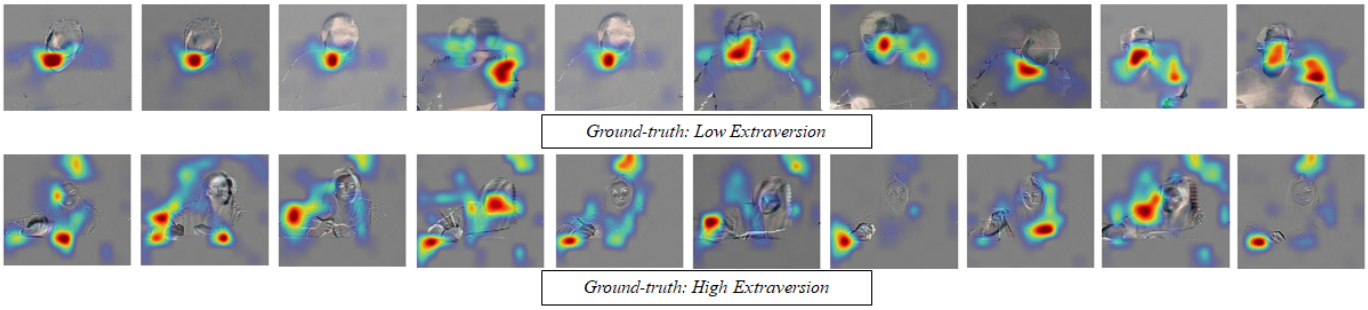

Fig. 2: For two example videos of ELEA-AV dataset [9] having different ground-truth labels, the visualization of the spatial saliencies overlaid with the top 10 key-dynamic images selected by the model. Red regions in the heat map representation are more discriminative than green regions, while regions without heat map means they are not discriminative for low/high extraversion classification task.

AlexNet+LSTM when dynamic images are the inputs to this model. AlexNet+LSTM allows us to detect the spatio-temporal saliency of each dynamic image [21], which is utilized to determine key-dynamic images. The deep VA-based NFs, extracted from key-dynamic images only, are combined using improved covariance-based feature encoding (I-CFE). This methodology was tested on automatic classification of individual personality traits, as low or high.

The experimental results showed that the proposed method performs much better than the state-of-the-art VA-based NFs. When it was tested on a meeting dataset (small group social interactions) [9], its performance is much better than all multi-modal approaches for all personality traits. Only for extraversion classification, it performs equally well with a multi-modal method using 73 different cues.

When the proposed method was evaluated on a more in-the-wild dataset composed of vlogs [24], it showed very promising results as compared to the baseline methods. It is important to recall that the baselines of vlog dataset [24] are either a multi-modal approach, or applying specific body parts detection and pre-processing steps such as face detection and alignment. More importantly, baselines were all customized to solve this specific dataset (see [24] for more information). Whereas, our method is based on VA-only, and it is more generic as not being based on specific body parts detection. We also do not apply any pre-processing step. In detail, we use a video frame as a whole, assuming that there are more foreground pixels than the background pixels (i.e. camera is close to the person) and we expect that dynamic image creation can discard the background. However, for vlog dataset [24], these assumptions fail when there is a camera motion or the camera is not close enough. Therefore, as future attempt, one can investigate whether pre-processing steps such as image cropping, person detection or background subtraction before creating dynamic images can positively contribute the performance of the proposed method or not.

Some further analyses were performed for low/high extraversion classification during meetings. In detail, the key-frame selection component of the proposed method showed better performance than *i)* the most popular key-frames selection algorithm in social/affective computing, *ii)* random and uniform dynamic images (or frames) selection, and *iii)* using all dynamic images (or frames). This is the first attempt that, deep learning-based spatio-temporal saliency detection is used for key-frames selection, and dynamic images are modeled with AlexNet+LSTM. The ablation study showed that *i)* using dynamic images instead of raw RGB images (i.e. frames) or MEIs performs better, *ii)* jointly modeling LSTM with CNN as compared to using CNN only has a positive contribution, *iii)* selecting key-dynamic images based on spatio-temporal saliency results in better performance as compared to using all dynamic images and *iv)* using the proposed I-CFE with SVM performs better than Softmax.

Better results of the proposed method is very important given that it is based on only one modality, which is in contrast to the majority of the baselines. It is highly possible that, much better classification results can be obtained when the proposed VA-based NFs are combined with audio-based and other video-based NFs. Besides, the performance of the proposed method can potentially be increased when it is combined with LMKL instead of SVM (like applied in [16], i.e. methodologically the most similar work to our work) or when ResNet is used instead of AlexNet (this can be more suitable once the size of the dataset is increased, given that ResNet is a deeper network).

One can approach the personality traits classification as a regression problem as well, but in this paper, we aim at classification, which requires binarization of the traits as low and high. This was performed in line with the baselines of each dataset. However, the proposed method can be adapted to process continuous valued annotations to perform regression as future work. A limitation of the proposed method is requiring the number of dynamic images to be selected as an input (similar to many other key-frames selection methods). In this study, the numbers used were enough to prove that the proposed method performs well, however, as further attempt; the number of key-dynamic images will be learned as a parameter during training. Creating dynamic images from optical flow images for the problem addressed in this study, is another issue to be investigated as future work. The e value (Eq. 3) included to the calculation of I-CFE, which resulted in better performances not only for the proposed method but also for the state-of-the-art [16], was set empirically, but it would be important to learn this value as a parameter during training. Another future investigation to be applied can be adapting attention-

based training [53] to improve the feature learning stage (AlexNet+LSTM).

Furthermore, the key-dynamic images selected can be used to obtain more insights regarding the task. For example, existing handcrafted VA-based NFs can be extracted from corresponding frames only and can be used for learning/classification.

The proposed method can potentially be applied to any application based on nonverbal behavior analysis, thanks to being data-driven (learning feature representations automatically from the training data itself). It can also be utilized to summarize the videos or to extract video segments such that the resulting much shorter videos can be used for annotation. In this way, annotation workload can be decreased a lot.

ACKNOWLEDGMENT

The authors thank Radoslaw Niewiadomski for the insightful comments and valuable suggestions, which helped a lot to improve the quality the paper.

REFERENCES

[1] K. Kalimeri, B. Lepri, T. Kim, F. Pianesi, and A. S. Pentland, "Automatic modeling of dominance effects using granger causality," *Human Behavior Understanding, Lecture Notes in Computer Science, A. Salah, Lepri, B., Eds., Springer, Berlin/Heidelberg*, vol. 7065, pp. 124–133, 2011.

[2] K. Kalimeri, B. Lepri, O. Aran, D. B. Jayagopi, D. Gatica-Perez, and F. Pianesi, "Modeling dominance effects on nonverbal behaviors using granger causality," in *Proceedings of ACM ICMI*, 2012, pp. 23–26.

[3] A. Sapru and H. Bourlard, "Automatic social role recognition in professional meetings using conditional random fields," in *Proceedings of Interspeech*, 2013.

[4] S. Favre, A. Dielmann, and A. Vinciarelli, "Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models," in *Proceedings of ACM MM*, 2009, pp. 585–588.

[5] S. Favre, H. Salamin, J. Dines, and A. Vinciarelli, "Role recognition in multiparty recordings using social affiliation networks and discrete distributions," in *Proceedings of ACM ICMI*, 2008, pp. 29–36.

[6] A. Vinciarelli, F. Valente, S. H. Yella, and A. Sapru, "Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the ami meeting corpus," in *Proceedings of the IEEE Int. Conf. on Systems, Man and Cybernetics*, 2011.

[7] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *Proceedings of ACM ICMI*, 2007, pp. 271–278.

[8] C. Beyan, N. Carissimi, F. Capozzi, S. Vascon, M. Bustreo, A. Pierro, C. Becchio, and V. Murino, "Detecting emergent leader in a meeting environment using nonverbal visual features only," in *Proceedings of ACM ICMI*, 2016, pp. 317–324.

[9] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 816–832, 2012.

[10] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Prediction of the leadership style of an emergent leader using audio and visual nonverbal features," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 441–456, 2018.

[11] B. Lepri, K. Kalimeri, and F. Pianesi, "Honest signals and their contribution to the automatic analysis of personality traits a comparative study," *Human Behavior Understanding, ser. Lecture Notes in Computer Science, A. Salah, T. Gevers, N. Sebe, and A. Vinciarelli, Eds. Springer, Berlin / Heidelberg*, vol. 6219, pp. 140–150, 2010.

[12] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proceedings of ACM ICMI*, 2008, pp. 53–60.

[13] O. Aran and D. Gatica-Perez, "One of a kind: inferring personality impressions in meetings," in *Proceedings of ACM ICMI*, 2013, pp. 9–13.

[14] S. Okada, O. Aran, and D. Gatica-Perez, "Personality trait classification via co-occurrent multiparty multimodal event discovery," in *Proceedings of ACM ICMI*, 2015, pp. 15–22.

[15] A. A. Kindiroglu, L. Akarun, and O. Aran, "Multi-domain and multi-task prediction of extraversion and leadership from meeting videos," *EURASIP Journal on Image and Video Processing*, vol. 77, pp. 1–14, 2017.

[16] C. Beyan, M. Shahid, and V. Murino, "Investigation of small group social interactions using deep visual activity-based nonverbal features," in *Proceedings of ACM Multimedia*, 2018, pp. 311–319.

[17] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001. [Online]. Available: http://dx.doi.org/10.1109/34.910878

[18] C. Beyan, V.-M. Katsageorgiou, and V. Murino, "Moving as a leader: Detecting emergent leadership in small groups using body pose," in *Proceedings of ACM Multimedia*, 2017, pp. 1425–1433.

[19] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proceedings of CVPR*, 2016.

[20] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of CVPR*, 2015.

[21] S. A. Bargal, A. Zunino, D. Kim, J. Zhang, V. Murino, and S. Sclaroff, "Excitation backprop for rnns," in *Proceedings of CVPR*, June 2018.

[22] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, 2017.

[23] R. Trabelsi, J. Varadarajan, Y. Pei, L. Zhang, I. Jabri, A. Bouallegue, and P. Moulin, "Multi-modal social interaction recognition using view-invariant features," in *Proceedings of ACM SIGCHI Inter. Workshop on Investigating Social Interactions with Artificial Agents*, 2017, pp. 47–48.

[24] V. Ponce-Lopez, B. Chen, M. Oliu, C. Cornearu, A. Clapes, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera, "ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results," in *European Conference on Computer Vision (ECCV 2016) Workshops*, Amsterdam, Netherlands, 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01381149

[25] P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona, "Large-scale isolated gesture recognition using convolutional neural networks," in *Proceedings of ICPR*, 2016.

[26] J. Wang, A. Cherian, and F. Porikli., "Ordered pooling of optical flow sequences for action recognition," in *Proceedings of IEEE WACV*, 2017.

[27] H. Wang, P. Wang, Z. Song, and W. Li, "Large-scale multimodal gesture segmentation and recognition based on convolutional neural networks," in *Proceedings of IEEE ICCV Workshops*, 2017.

[28] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Trans. Affective Computing*, pp. 1–1, 2017.

[29] S. Xiang, W. Rong, Z. Xiong, M. Gao, and Q. Xiong, "Visual and audio aware bi-modal video emotion recognition," in *Proceedings of CogSci*, 2017.

[30] Y. Baveye, E. Dellandra, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *Proceedings of Affective Computing and Intelligent Interaction*, 2015.

[31] Y. Zhu, Z. Jiang, J. Peng, and S. hua Zhong, "Video affective content analysis based on protagonist via convolutional neural network," in *Proceedings of 17th Pacific-Rim Conf. on Multimedia*, 2016, pp. 170–180.

[32] S. Zhalehpour, Z. Akhtar, and C. E. Erdem, "Multimodal emotion recognition with automatic peak frame selection," in *Proceedings of 2014 IEEE Inter. Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, 2014, pp. 116–121.

[33] F. Dornaika and B. Raducanu, "Efficient facial expression recognition for human robot interaction," *Computational and Ambient Intelligence*, pp. 700–708, 2007.

[34] Y. wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.
[35] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, pp. 79–82, 2007.
[36] F. Alam, E. A. Stepanov, and G. Riccardi, "Personality traits recognition on social network-facebook," in *Proceedings of Inter. AAAI Conf. on Weblogs and Social Media*, 2013.
[37] L. Qiu, H. Lin, J. Ramsay, and F. Yang, "You are what you tweet: Personality expression and perception on twitter," *Journal of Research in Personality*, vol. 46, no. 6, pp. 710–718, 2012.
[38] J.-I. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *IEEE Trans Multimedia*, vol. 15, pp. 41–55, 2012.
[39] 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 Trans. Affective Computing*, vol. 6, pp. 193–205, 2015.
[40] G. Mohammadi, A. Origlia, M. Filippone, and A. Vinciarelli, "From speech to personality: mapping voice quality and intonation into personality differences," in *Proceedings of ACM Multimedia*, 2012, pp. 789–792.
[41] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Mining large-scale smartphone data for personality studies," *Personal and Ubiquitous Computing*, vol. 17, no. 3, pp. 433–450, 2013.
[42] S. Okada, L. S. Nguyen, O. Aran, and D. Gatica-Perez, "Modeling dyadic and group impressions with inter-modal and inter-person features," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2018.
[43] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion-a systematic study," *IEEE Transactions on Affective Computing*, vol. 3, pp. 443–455, 2012.
[44] F. Valente, S. Kim, and P. Motlicek, "Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus," in *Proceedings of INTERSPEECH*, 2012, pp. 1183–1186.
[45] B. Aydin, A. A. Kindiroglu, O. Aran, and L. Akarun, "Automatic personality prediction from audiovisual data using random forest regression," in *Proceedings of ICPR*, 2016.
[46] Y.-S. Lin and C.-C. Lee, "Using interlocutor-modulated attention blstm to predict personality traits in small group interaction," in *Proceedings of ACM ICMI*, 2016, pp. 163–169.
[47] J. Staiano, B. Lepri, R. Subramanian, N. Sebe, and F. Pianesi, "Automatic modeling of personality states in small group interactions," in *Proceedings of ACM Multimedia*, 2011, pp. 989–992.
[48] O. Aran and D. Gatica-Perez, "Cross-domain personality prediction: from video blogs to small group meetings," in *Proceedings of ACM ICMI*, 2013, pp. 127–130.
[49] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Employing social gaze and speaking activity for automatic determination of the extraversion trait," in *Proceedings of ACM ICMI-MLMI*, 2013.
[50] K. Kalimeri, B. Lepri, and F. Pianesi, "Causal-modelling of personality traits: extraversion and locus of control," in *2nd Inter. Workshop on Social Signal Processing*, 2010.
[51] S. Fang, C. Achard, and S. Dubuisson, "Personality classification and behaviour interpretation: an approach based on feature categories," in *Proceedings of ACM ICMI*, 2016, pp. 225–232.
[52] J. Staiano, B. Lepri, K. Kalimeri, N. Sebe, and F. Pianesi, "Contextual modeling of personality states? dynamics in face-to-face interactions," in *Proceedings of IEEE Inter. Conf. on Privacy, Security, Risk, and Trust, and IEEE Inter. Conf. on Social Computing*, 2011, pp. 898–899.
[53] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of ICCV*, 2017, pp. 3456–3465.

Cigdem Beyan received her MSc. degree in Informatics from Middle East Technical University, Ankara, Turkey in 2010 and her Ph.D. degree in Informatics from University of Edinburgh, Edinburgh, UK in 2015. She is currently a Postdoctoral Researcher at the Istituto Italiano di Tecnologia, Genoa, Italy in the department of Pattern Analysis and Computer Vision. She has coauthored over 30 papers published in refereed journals and international conferences. Among her main research interest, there are social signal processing, multi-modal data analysis, human/animal behavior understanding, deep learning and classification of imbalanced data. She is a reviewer of most significant multimedia, affective computing, computer vision and pattern recognition journals including multiple IEEE Transactions, and IEEE/ACM conferences. She is a Guest Co-Editor of a special issue in Frontiers in Robotics and AI and in the Editorial Board of ICES Journal of Marine Science covering area of applications of computer vision and machine learning in marine science. She is a member of IEEE since 2013 and an Associate Fellow of the Higher Education Academy UK since 2014.

Andrea Zunino received his M.Sc. degree in Multimedia Signal Processing and Telecommunication Networks from the University of Genoa (Italy) in 2014. He pursued a PhD degree in Science and Technology for Electronic and Telecommunication Engineering at Pattern Analysis and Computer Vision department, Istituto Italiano di Tecnologia in 2018. Currently, he is a Postdoctoral Researcher at the same department. His research activities focus on machine learning and computer vision with particular interests in audio/video signals processing, activity recognition/prediction for video-surveillance and psychological intention-understanding, interpretable and explainable deep learning.

Muhammad Shahid received his M.S. degree (2016) in Computer Engineering with major in computer vision from Pakistan. He is currently pursuing his Ph.D in Istituto Italiano di Tecnologia, Genoa, Italy in the department of Pattern Analysis and Computer Vision with the collaboration of Department of Naval, Electrical, Electronic and Telecommunications Engineering from University of Genoa, Italy. His current research includes nonverbal communication, social signal processing, multimodal data fusion and deep learning.

Vittorio Murino is currently full professor at the University of Verona, Italy, and a senior video intelligence expert at Huawei Technologies in Dublin, Ireland. He took his Ph.D. degree in Electronic Engineering and Computer Science in 1993 at the University of Genova, Italy. In 1995-1998, he was assistant professor at the Dept. of Mathematics and Computer Science of the University of Udine, Italy, and since 1998 he has been working at the University of Verona. He was the chairman of Department of Computer Science from 2001 to 2007, and the coordinator of Ph.D. program in Computer Science from 1999 to 2003. He is a scientific responsible of several national and European projects, and evaluator of EU project proposals related to several frameworks and programs. From 2009 to 2019, he worked at the Istituto Italiano di Tecnologia in Genova, Italy, as the director of Pattern Analysis and Computer Vision department. His main research interests include computer vision, pattern recognition and machine learning, more specifically, techniques for image and video processing for (human) behavior analysis and related applications such as video surveillance, and biomedical imaging. He is the co-author of more than 400 papers published in refereed journals and international conferences, member of the technical committees of important conferences (CVPR, ICCV, ECCV, ICPR, ICIP, etc.), and is guest co-editor of special issues in relevant scientific journals. He is member of the editorial board of Computer Vision and Image Understanding, Machine Vision & Applications, and Pattern Analysis and Applications journals. He is a senior member of the IEEE and a fellow of the IAPR.