Deep Learning-based Assessment of Facial Periodic Affect in Work-like Settings

Siyang Song[1] *, Yiming Luo[2] *, Vincenzo Ronca[3], Gianluca Borghini[3], Hesam Sagha[4], Vera Rick[5], Alexander Mertens[5], and Hatice Gunes[1]

[1] AFAR Lab, University of Cambridge, United Kingdom
[2] Tsinghua University, China
[3] BrainSigns, Italy
[4] audEERING, Germany
[5] RWTH Aachen University, Germany

Abstract. Facial behaviour forms an important cue for understanding human affect. While a large number of existing approaches successfully recognize affect from facial behaviours occurring in daily life, facial behaviours displayed in work settings have not been investigated to a great extent. This paper presents the first study that systematically investigates the influence of spatial and temporal facial behaviours on human affect recognition in work-like settings. We first introduce a new multi-site data collection protocol for acquiring human behavioural data under various simulated working conditions. Then, we propose a deep learning-based framework that leverages both spatio-temporal facial behavioural cues and background information for workers' affect recognition. We conduct extensive experiments to evaluate the impact of spatial, temporal and contextual information for models that learn to recognize affect in work-like settings. Our experimental results show that (i) workers' affective states can be inferred from their facial behaviours; (ii) models pre-trained on naturalistic datasets prove useful for predicting affect from facial behaviours in work-like settings; and (iii) task type and task setting influence the affect recognition performance.

1 Introduction

Human affect is a key indicator of various human internal states including mental well-being [3, 17] and personality [14, 12], as well as people's working behaviours [33, 13, 11]. Therefore, accurately understanding employees' affect in their working environment would enable managers identify risks for workers' safety, collect a history of risk exposures and monitor individuals' health status, which would further help employers in shaping organizational attitudes and decisions.

Since facial behaviours are a reliable source for affect recognition and can be easily recorded in an non-invasive way, a large number of existing approaches have been devoted to inferring affect from the human face. These approaches frequently claim that static facial displays, spatio-temporal facial behaviours

* Equal contribution.

and even the background in a face image can provide useful cues for affect recognition. Subsequently, existing approaches can be categorized as: static face-based solutions that infer affect from each static facial display [24, 10, 7], full frame-based solutions that utilize not only facial display but also background cues [1], and spatio-temporal solutions that consider facial temporal evolution of the target face [15, 28, 22, 5].

Although some of these approaches can accurately identify human affect in both in-lab and naturalistic conditions, none of these works have focused on analysing affect in working environments. This is largely due to the fact that there is neither a well-developed protocol for worker facial behaviour data collection nor a publicly available worker facial behaviour dataset for developing affect recognition systems in work settings (**Problem 1**). Since different working conditions would lead workers to express behaviours in different manners [6, 18, 29], existing approaches that are developed using naturalistic or in-lab facial data may fail to accurately recognize affect from facial expressions expressed in work environments (**Problem 2**). Another key issue is that these approaches mainly provide affect prediction for a single facial image or video frame. However, in real-world working conditions, a large number of workers may need to be recorded and assessed many times a day. Subsequently, the limited computing and disk resources that are typical for SMEs (small-to-medium scale businesses) would not always allow making frame-level affect predictions in real time, and storing these for a high number of workers (**Problem 3**). More specifically, a common and realistic requirement for a real-world worker affect recognition system that could be adopted by SMEs is to predict each worker's affective state regularly but for a certain period of time, i.e., providing periodical affect predictions.

In this paper, we aim to address the three problems described above. Firstly, we introduce a new data collection protocol for acquiring naturalistic human audio-visual and physiological behavioural signals stimulated under different work-like conditions, tasks and stress levels. Using this protocol, we acquired the first human working facial behaviour database called WorkingAge DB, which is collected in four different sites with participants of different backgrounds (**addressing Problem 1**). Then, we benchmark several standard deep learning-based solutions on the collected WorkingAge DB, providing a set of baseline results for worker's periodical facial affect recognition (**addressing Problem 2 and Problem 3**). We further investigate the influence of the frame-rate on different models' periodical facial affect recognition performance with the assumption that SMEs where such a system is to be deployed typically have limited computational and disk resources, as well as the influence of task type, recording site, gender, and various feature representations. In summary, the main contributions of this paper are listed as follows:

- We propose an new protocol to acquire human facial behavioural data under various simulated working conditions together with the self-reported emotion/affect/workload state for each condition.
- Based on the proposed protocol, we collect a facial behaviour dataset in work-like settings across multiple sites with participants from different cultural

and language backgrounds. To the best of our knowledge, this is the first cross-cultural human facial behaviour dataset that is collected under various simulated working conditions.
– We benchmark several standard deep learning-based video-level behaviour approaches for worker's periodical dimensional affect recognition, and specifically investigate the influence of influence of task type, recording site, gender, and feature representations on models' performance. This provides a set of baselines for future studies that will focus on facial behaviour understanding in work settings. **Code access:** our code for the experiments is made available at https://github.com/takuyara/Working-Age-Baselines.

2 A protocol for human facial behaviour data acquisition in work-like settings

Although many existing face datasets [21, 20, 15, 16, 23, 32] have been annotated with dimensional affect labels, to the best of our knowledge, none of them has been collected in a working environment. Moreover, these datasets only provided static face/frame-level labels without periodical affect labels (labels that reflect the subject's affective state for a certain period of time (each clip in this paper)). As a result, none of them is suitable to be used for the purpose of investigating the relationship between human facial behaviours and affective states in working environments. To bridge this research gap, we propose a new protocol for acquiring human facial behavioural data displayed in various simulated working conditions, and annotated with self-reported affect labels that reflect workers' periodical affective states. The details of this protocol is described below.

Sensors setup: The sensor setup of the protocol is illustrated in Fig. 1. During the recording, the participant sits at a table, where a laptop is placed to display slides that guide the participant to undertake a number of tasks based on a pre-defined order. To record visual information (including facial behaviours), a Logit web camera is placed in front of the participant. Additionally, a GoPro camera is also placed on the keyboard of the laptop to record facial behaviours during the Operations Task Game when the participant lowers the head on a panel to pick up items. Specifically, we build our model based on facial behaviours recorded by both cameras, i.e., only facial behaviours triggered by the operation task (explained in following paragraphs) are recorded by the GoPro camera.

Work-like tasks To simulate several working conditions, the proposed protocol consists of three work-like tasks including the N-back tasks, the video conference tasks and the operation game (i.e., the Doctor game). Additionally, we set an Eyes Open and Close task as the first task to acquire the baseline behaviours of each participant. The details of the four tasks are listed below:

– **Eyes Open and Close:** this task contains two sub-tasks, **(i) Eyes Open**: keeping the eyes open for 1 minute; and **(ii) Eyes Closed**: keeping the eyes closed for 1 minute. This task aims to acquire the baseline behaviours of each participant and help them get familiar with the task and the data acquisition environment.

Fig. 1. Hardware settings and example recordings, where the bottom-left figure displays the participant conducting a sub-task under the 'stressful condition' and is recorded by the logit camera. The bottom-right figure displays the participant conducting an operation sub-task and is recorded by the GoPro camera.

- **N-back task:** this task is a continuous performance task that is commonly employed to simulate different working memory capacity [9]. In our protocol, it simulates an office-related activity that does not need intensive physical work but causes mental strain. This task contains six sub-tasks in the following order: **(i) Baseline (NBB)**: looking at the N-back interface for 1 minute without reacting; **(ii) Easy game 1 (NBE01)**: playing 0-back game for 2 minutes (typing the number that has just been shown on the screen); **(iii) Hard game 1 (NBH01)**: playing 2-back game for 2 minutes (typing the number that has just been shown on the screen two turns ago); **(iv) Hard game 2 (NBH02)**: playing 2-back game for 2 minutes; **(v) Easy game 2 (NBE02)**: playing 0-back game for 2 minutes; and **(vi) Stressful hard game (NBS)**: playing 2-back game for 2 minutes with 85 dB background noise while a human experimenter is seated in the same room as the participant.
- **Video conference task:** this task simulates a teleworking scenario, in which employees are frequently requested to interact and coordinate with colleagues who are not physically present. It contains three sub-tasks: **(i) Baseline (WEB)**: looking at Microsoft Teams screen without reacting; **(ii) Positive emotions (WEP)**: describing the happiest memory in one's life to the human experimenter for 2 minutes via Microsoft Teams video conferencing application (aiming to stimulate positive emotions); and **(iii) Negative**

emotions (WEN): describing the most negative/sad memory in one's life to the human experimenter for 2 minutes via Microsoft Teams video conferencing application (aiming to stimulate negative emotions). During both tasks, if needed, the human experimenter would ask several neutral and factual questions to keep the participant talking for about 2 minutes.

- **Operation task (Doctor Game):** this task requires the participant to use tweezers to pick up objects from a panel, simulating an assembly line scenario. It contains six sub-tasks: **(i) Baseline (DB)**: looking at the doctor game panel without reacting; **(ii) Easy game 1 (DE01)**: picking up and removing 5 objects from the panel within 2 minutes; **(iii) Hard game 1 (DH01)**: picking up and removing as many objects as possible from the panel within 3 minutes; **(iv) Easy game 2 (DE02)**: picking up and removing 5 objects from the panel within 2 minutes; **(v) Hard game 2 (DH02)**: picking up and removing as many objects as possible from the panel within 3 minutes; and **(vi) Stressful hard game (DS)**: removing as many objects as possible from the panel within 3 minutes with 85 dB background noise while a human experimenter is seated in the same room as the participant.

After finishing each sub-task, the participant is asked to fill in two questionnaires.

Self-reported questionnaires In our protocol, we propose to obtain periodical emotion and affect annotations from each participant, i.e., sub-task-level emotion and affect annotations. Specifically, two questionnaires are employed, namely, Geneva Emotion Wheel (GEW) [25] that measures the intensities of several categorical emotions of the participant along 5 scales (from low to high), and Self-Assessment Manikin (SAM) [4] that measures the intensities of three affect dimensions (arousal, valence and power) from unhappy/calm/controlled (1) to happy/excited/in-control (9) using a scale of 9.

Data acquisition The study was approved by the relevant Departmental Ethics Committee. Additional COVID-19 related measures were also put in place prior to the study. Prior to data acquisition, each participant is provided with an information sheet and is asked to read and sign a consent form. Following these procedures, the experimenter explains all the details (e.g., purpose, tasks, questionnaires, etc.) of the study to the participant. Then, the participant is asked to enter the room and sit in front of the laptop. The experimenter then leaves the recording the room and goes to the operations room next door to remotely start both cameras. The instructions related to each work-like task are displayed to the participant on the laptop screen. After each main task (N-back, video conference and operation tasks), the participant is instructed to relax for 4 minutes by listening a calming music. This aims to help the participant to get back to their baseline / neutral affective and cognitive state, to prevent the positive or negative emotions caused by the previous tasks from impacting the future tasks.

3 The WorkingAge Facial Behaviour Dataset

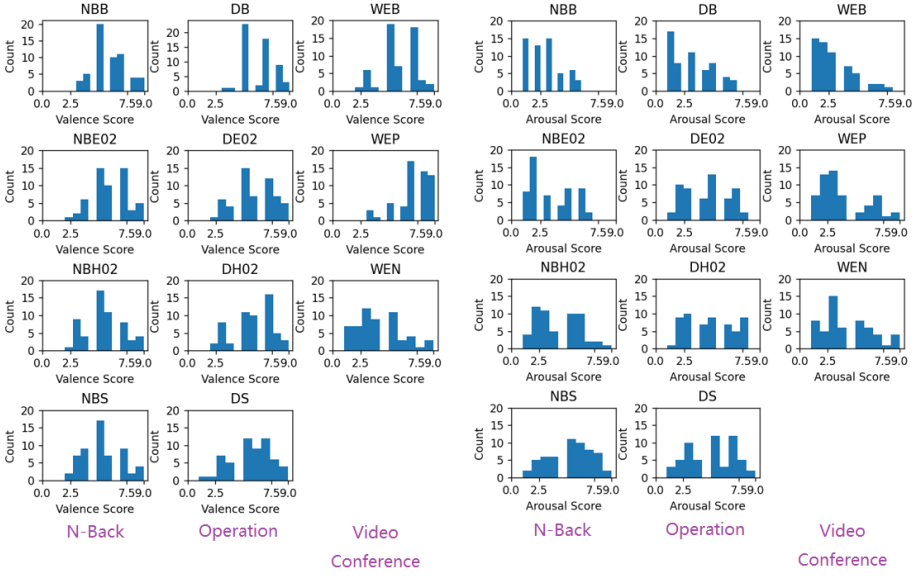
Based on the proposed protocol, we acquired a human facial behaviour dataset in work-like settings in four different sites located in three countries: the audEERING and RWTH Aachen University in Germany, the University of Cambridge in United Kingdom, and the BrainSigns in Italy (i.e., they are individually referred as AUD, RWTH, UCAM and BS in this paper). The collected dataset contains data from a total of 55 participants, with 7 participants coming from the AUD site, 16 participants coming from the BS site, 20 participants coming from the RWTH site, and 12 participants coming from the UCAM site. It contains 935 clips (each clip corresponds to a sub-task, and 17 sub-tasks were recorded for each participant), where 605 of them are annotated using the proposed protocol (i.e., participants were not asked to provide self report annotations for NBE01, NBH01, DB01 and DE01 tasks to reduce their annotation burden. As a result, these clips were not used for experiments.). Specifically, the videos collected by Logit camera and GoPro Camera are set as 24 and 30 fps during the recording, with the frame resolutions of 1280×720, and 1920×1080, respectively. The mean value, standard deviation, maximum and minimum values of each annotated sub-task's duration in our dataset are listed in Table. 1. It is clear that the participants need longer periods of time to undertake the WEP and WEN tasks, while DB, DE02 and DS tasks take shorter time to complete.

Sub-task	Mean	Standard Deviation	Maximum	Minimum
NBB	73.1	15.9	111.4	25.9
NBE02	91.4	29.8	131.2	26.1
NBH02	93.4	29.8	135.3	29.1
NBS	97.1	27.0	129.9	36.5
WEB	57.8	13.5	81.2	20.8
WEP	136.6	40.2	206.2	49.0
WEN	134.5	35.0	224.0	54.1
DB	61.0	7.2	77.5	38.1
DE02	60.6	41.2	204.2	10.5
DH02	100.3	41.8	185.3	32.4
DS	61.6	39.3	212.1	14.2

Table 1. Statistics of clips' duration (seconds) in the WorkingAge dataset.

In addition, the distributions of the self-reported arousal and valence labels (based on the SAM questionnaire) are illustrated in Fig. 2. In this paper, we further quantize the valence/arousal labels into three classes: positive (scores 7, 8, or 9), neutral (scores 4, 5, or 6) and negative (scores 1, 2, or 3). As we can see, during the three baseline sub-tasks, participants reported a relatively high valence (72, 88 and 11 subjects have positive, neutral and negative valence status) and low arousal value (6, 46 and 119 subjects have positive, neutral and negative arousal). However, it can be seen that the increasing mental workload

requirement clearly leads participants to report lower valence (with mean valence scores of 5.9, 5.8 and 5.7 for DE02, DH02 and DS as well as 6.0, 5.5 and 5.3 for NBE02, NBH02 and NBS) and higher arousal (with mean arousal scores of 4.4, 4.8 and 5.0 for DE02, DH02 and DS as well as 3.4, 4.0 and 5.2 for NBE02, NBH02 and NBS). We also see that the valence label distributions are quite different for the three video conference sub-tasks, i.e., most participants reported a relatively neutral valence status during the baseline condition, while positive and negative memory-based conversations caused most participants to report corresponding positive and negative valence values, with mean valence scores of 5.8, 7.1 and 4.0 for WEB, WEP and WEN, respectively.



a). The distribution of self-reported valence intensities for each subtask

b). The distribution of self-reported arousal intensities for each subtask.

Fig. 2. The distribution of the self-reported valence intensities for each subtask.

4 Periodical facial affect recognition

Although facial behaviours have been proven to be informative for inferring human affect under natural or controlled conditions, none of the previous studies evaluated the feasibility of using workers' facial behaviours to infer affect in terms of valence and arousal. In this section, we implement three standard deep learning-based short video-level modelling approaches to provide a benchmark for the task of facial affect recognition in work-like settings.

Baseline 1 Given a short face video, the first baseline starts with generating frame-level affect predictions for all frames, which are then combined to output periodical affect prediction. In particular, we individually employed two frame-level facial analysis models, i.e., a ResNet-50 [8] that is pre-trained for facial expression recognition (i.e., pre-trained on the the FER 2013 dataset) and a GraphAU model [19] that is pre-trained for facial action units (AUs) recognition (i.e., pre-trained on the BP4D dataset [31]). Specifically, we individually use the latent feature output by the second-last fully connected layer of the ResNet-50, as well as the 12 AU predictions generated by the GraphAU model, as the frame-level facial features. Then, we individually apply a multi-layer perceptron (MLP) on each of them to provide frame-level valence and arousal predictions, which is trained by re-using the clip-level self-reported valence/arousal scores as the frame-level label. To obtain periodical (clip-level) affect predictions, we combine all frame-level predictions of the target clip with the following widely-used strategies: (i) using the mode prediction of all frame-level predictions as the periodical affect predictions (i.e., GraphAU(P)-MODE and ResNet(P)-MODE); (ii) applying a Long-short-term-memory Network (LSTM) to combine all frame-level predictions (i.e., GraphAU(P)-LSTM and ResNet(P)-LSTM); and (iii) applying spectral encoding algorithm [26, 27] to produce a spectral heatmap from all frame-level predictions, which is then fed to a 1D-CNN to generate periodical affect predictions (i.e., GraphAU(P)-SE and ResNet(P)-SE).

Baseline 2 The second baseline also applies the same two pre-trained models used in baseline 1 to provide frame-level facial features. Differently from the baseline 1, we employ three long-term modelling strategies to combine all frame-level facial features (the latent feature vectors produced by the last FC layer of the corresponding frame-level model) of the clip as the clip-level (periodical) affect representation: (i) averaging all frame-level facial features (i.e., GraphAU(F)-AVERAGE and ResNet(F)-AVERAGE); (ii) applying LSTM to process all frame-level facial features (i.e., GraphAU(F)-LSTM and ResNet(F)-LSTM); and (iii) spectral encoding all frame-level facial features (i.e., GraphAU(F)-SE and ResNet(F)-SE). These clip-level affect representations are then fed to either an MLP (for (i)) or 1D-CNN (for (ii) and (iii)) to generate clip-level affect predictions.

Baseline 3 The third baseline applies a spatio-temporal CNN (Temporal Pyramid Network (TPN) [30]) to process the facial sequence. In particular, we first divide each clip into several segments, where each consists of 160 frames, and down-sample each segment to 32 frames. We then feed the cropped face sequence (32 frames) to TPN for affect classification. If a clip contains multiple segments, then the clip-level predictions are achieved by averaging all segment-level predictions.

5 Experiments

5.1 Experimental setup

Data pre-processing: We re-sampled all video clips to 24 fps to make videos recorded by different cameras to have the same frame rate. Then, we used OpenFace 2.0 [2] to crop and align the face from each frame, where frames with failed or low-confidence face detection are treated as black images.

Training and evaluation protocol: We use the leave-one-site-out validation protocol for models' training and evaluation, i.e., at each time, we use all clips from three sites to train the model, and evaluate the trained model on the rest one. The final reported results are obtained by averaging validation results of four folds.

Model settings and training details: There are three main settings for our baseline models: (i) frame-level feature extraction; (ii) clip-level (periodical) facial behaviour representation extraction; and (iii) classifiers. Specifically, we used the output (2048D) of the second-last layer of the ResNet-50 and the 12 predicted action units' occurrences as the frame-level facial features, respectively. In terms of the clip-level representation extraction, we used bidirectional LSTM with one hidden layer, as well as a spectral encoding algorithm with a resolution of 256 (80 lowest frequencies are selected). Finally, the MLP classifier is set as 3 layers for mode/averaging-based classification while the 1D-CNN is set to have 3 convolution blocks where each consists of a convolution layer, a ReLU activation, as well as a dropout layer, whose channels and hidden sizes varies depending the size of the input feature. We used the batch size of 512 and initial learning rate of 0.001 for all experiments.

Evaluation metrics: Considering that the samples are unbalanced in terms of arousal and valence label distibution, in this paper we employ the Unweighted Average Recall (UAR) as the measurement to evaluate different baseline performances on facial affect recognition in work-like settings.

5.2 Baseline results of leave-one-site-out cross-validation

Table 2 and Table 3 list the valence and arousal classification UAR results achieved by all baseline systems for all sub-tasks (the models are trained using all clips in the training set regardless of the task type). It can be seen that almost all baselines achieved over the chance-level classification UAR (33.33%), with the *GraphAU(P)-SE* system achieving the best valence UAR result (42.31%) and the *ResNet(F)-SE* system achieving the best arousal UAR result (41.20%). Meanwhile, we found that if we only use the mode prediction of all frame-level predictions, both valence and arousal classification results are clearly worse than most of other systems, i.e., the two corresponding systems only achieved less than 34% valence and arousal classification accuracy. These results indicate that: (i) according to Fig. 3, the long-term modelling for either frame-level predictions or features is a crucial step to achieve more reliable periodical arousal/valence

Model	NBB	NBE02	NBH02	NBS	DB	DE02	DH02	DS	WEB	WEP	WEN	Total
GraphAU(P)-SE	0.3814	0.4279	0.4206	0.4691	0.6667	**0.4026**	0.4611	**0.4461**	0.5324	0.3280	0.2857	**0.4231**
GraphAU(P)-LSTM	0.3921	0.4254	0.4444	0.4414	0.6667	0.3898	0.4611	0.4428	0.5139	0.2899	0.3175	0.4209
GraphAU(P)-MODE	0.2918	0.2667	0.3254	0.3241	0.5417	0.2700	0.2019	0.2525	0.3287	0.2899	0.2619	0.3009
GraphAU(F)-SE	0.3584	0.4019	0.4246	**0.4784**	0.6875	0.3929	0.3963	0.3956	0.4491	0.2984	0.3810	0.3995
GraphAU(F)-LSTM	0.3921	0.4254	0.4444	0.4414	0.6667	0.3898	0.4611	0.4428	0.5139	0.3090	0.2857	0.4153
GraphAU(F)-AVERAGE	0.3685	0.3994	0.3651	0.3951	0.6250	0.3570	0.4500	0.4209	0.4722	**0.3640**	0.3333	0.3970
ResNet(P)-SE	0.3653	0.4402	0.4243	0.4484	0.7150	0.3796	0.4956	0.3881	0.5207	0.3564	0.3254	0.4226
ResNet(P)-LSTM	0.3847	**0.4438**	0.4473	0.4444	0.6558	0.3977	0.4974	0.4171	0.5207	0.3023	0.2857	0.4176
ResNet(P)-MODE	0.3153	0.3394	0.3798	0.2897	0.4558	0.2778	0.3465	0.3124	0.3622	0.3504	**0.4127**	0.3344
ResNet(F)-SE	**0.3951**	**0.4438**	**0.4473**	0.4722	0.6550	0.3750	**0.5140**	0.4211	**0.5393**	0.3023	0.2857	0.4225
ResNet(F)-LSTM	0.3847	**0.4438**	0.4358	0.4444	0.6542	0.3838	**0.5140**	0.4171	**0.5393**	0.3023	0.2857	0.4191
ResNet(F)-AVERAGE	0.3847	**0.4438**	**0.4473**	0.4444	**0.7167**	0.3801	0.4974	0.4316	0.5207	0.3023	0.2857	0.4219
TPN	0.3751	0.3740	0.3016	0.2940	0.4785	0.2686	0.2571	0.4002	0.3062	0.2915	0.3379	0.3350

Table 2. The UAR results achieved for worker's valence recognition, where the name of each method is formatted as *frame level facial feature-long term model*, where P and F represent the frame-level prediction and facial features, respectively. For example, *ResNet(P)-SE* denote the system that applies ResNet facial features to make frame-level affect predictions, and then using spectral encoding algorithm to summarise all frame-level valence/arousal predictions as the clip-level valence/arousal prediction.

Model	NBB	NBE02	NBH02	NBS	DB	DE02	DH02	DS	WEB	WEP	WEN	Total
GraphAU(P)-SE	0.5605	0.3417	0.3506	0.4058	0.3464	**0.4842**	0.4077	**0.4118**	0.3538	0.4167	0.3801	0.3999
GraphAU(P)-LSTM	0.5474	0.3625	0.3975	0.4058	0.4071	0.3667	0.3310	0.3725	0.3547	0.3774	0.3581	0.3678
GraphAU(P)-MODE	0.5658	**0.4319**	0.3232	0.2580	0.2389	0.3741	0.3902	0.3081	0.3155	**0.4358**	0.2327	0.3427
GraphAU(F)-SE	0.5711	0.3819	**0.4149**	0.4014	0.4224	0.4201	0.3634	0.3880	0.4153	0.4100	0.3973	0.3967
GraphAU(F)-LSTM	0.5474	0.4153	0.4134	0.4058	0.4071	0.3667	0.3310	0.3725	0.3645	0.3774	0.3581	0.3759
GraphAU(F)-AVERAGE	0.5974	0.4042	0.3983	0.4203	**0.4309**	0.3667	0.3310	0.3725	0.3645	0.3774	0.3581	0.3790
ResNet(P)-SE	0.5244	0.3636	0.3837	0.3610	0.4115	0.3842	0.3860	0.3889	0.3873	0.3988	**0.4074**	0.3834
ResNet(P)-LSTM	0.5231	0.3206	0.3692	0.3900	0.4122	0.3491	0.3651	0.4074	0.3775	0.4129	0.3454	0.3671
ResNet(P)-MODE	0.4359	0.3011	0.3202	0.3320	0.3118	0.3667	0.3684	0.3519	0.3595	0.4014	0.3639	0.3386
ResNet(F)-SE	**0.6346**	0.4133	0.3775	0.4295	0.4036	0.4719	0.4106	0.3519	0.4101	0.3775	0.3406	**0.4120**
ResNet(F)-LSTM	0.5462	0.3925	0.4037	**0.4466**	0.4029	0.4649	**0.4496**	0.4021	**0.4869**	0.3837	0.3285	0.4032
ResNet(F)-AVERAGE	0.5231	0.3945	0.3996	0.3900	0.4122	0.3991	0.4002	0.4074	0.3954	0.3758	0.3639	0.3858
TPN	0.4823	0.4877	0.3543	0.3403	0.3168	0.1935	0.2391	0.2951	0.1807	0.2602	0.3498	0.3182

Table 3. The UAR results achieved for worker's arousal recognition.

predictions, as simply choosing the mode prediction from all frame-level predictions or averaging all frame-level features clearly provided the worst results; (ii) the frame-level facial analysis (AU recognition/facial expression recognition) models that are pre-trained using the lab-based facial datasets (AU or facial expression datasets) can still extract human affect-informative facial features from facial displays triggered by work-like tasks, as their features frequently provide around 40% UAR for both tasks (the chance-level UAR should be around 33% for three class classification); and (iii) directly pairing workers' facial sequences with clip-level affect labels to train spatio-temporal models does not provide superior results, which further validates that frame-level facial analysis models pre-trained on facial datasets acquired in naturalistic settings are beneficial for facial affect analysis in work-like settings.

5.3 Ablation studies

Task type: Since workers' facial behaviours are highly correlated with the task type and task setting, we also specifically investigate which tasks can trigger most affect-informative facial behaviours in Fig. 4, where we report the average

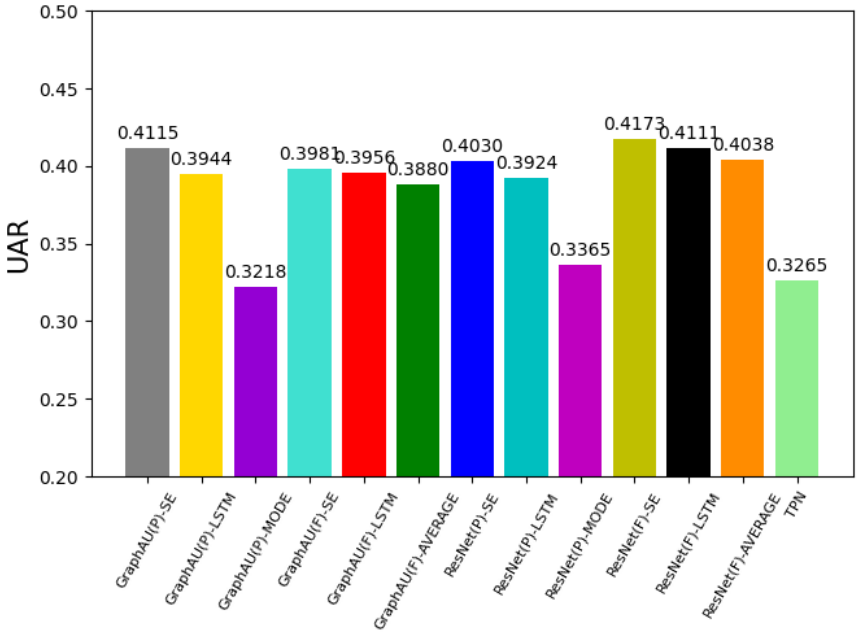


Fig. 3. The average valence and arousal UAR results of all baseline models.

UAR results achieved by all baselines for each task. The facial behaviours triggered by four sub-tasks allow the model to achieve over 40% valence recognition UAR results, which are clearly superior than the results achieved on other sub-tasks. Therefore, we hypothesize that different subjects display affect in different intensity (valence) when undertaking different sub-tasks of the same task (even though their facial behaviour may be similar). Meanwhile, facial behaviours displayed during N-back baseline sub-task is very informative for predicting subjects' arousal, i.e., the arousal UAR result achieved for the baseline sub-tasks of N-Back has more than 14.76% absolute accuracy improvements over the results of other sub-tasks. This finding suggests that human facial displays before conducting the memory task may be reliable for inferring worker's arousal state.

Validation set	AUD	BS	RWTH	UCAM
Valence	**0.3715**	0.3319	0.3438	0.3337
Arousal	**0.3621**	0.3385	0.3405	0.3523

Table 4. The results of the four-fold cross-validation results achieved by our best models (GraphAU(P)-SE for valence, ResNet(P)-SE for arousal).

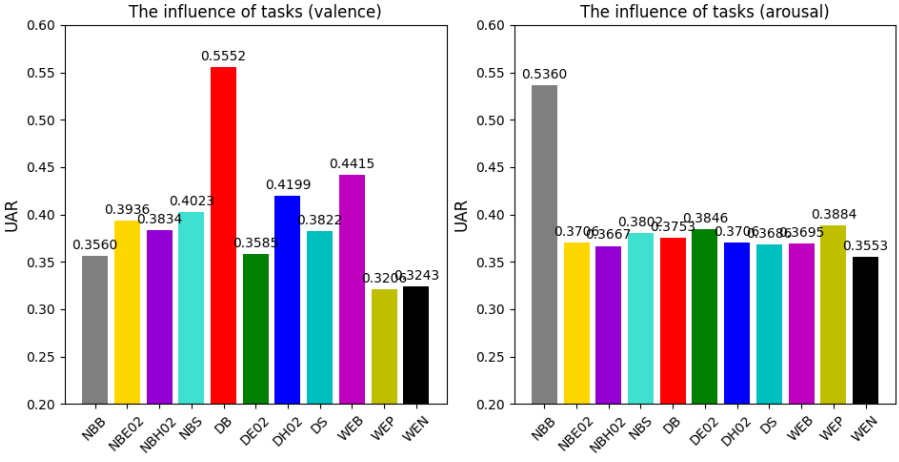


Fig. 4. The influence of different tasks on valence and arousal prediction.

Recording site: We also explore the differences in affect classification for different sites. Table 4 displays the leave-one-site-out four-fold cross-validation results. It is clear that the data collected at different sites impact the valence classification results, with around 4% UAR difference between the lowest (0.3319 (BS)) and the highest (0.3715 (AUD)). These results indicate that people at different sites may display different facial behaviours when expressing valence. On the other hand, the performance variations for arousal classification are much smaller, indicating that the relationship between arousal and workers' facial behaviours are more stable as compared to valence.

Gender: We report the gender-dependent worker valence/arousal classification UAR results using our all baselines, where leave-one-site-out cross validation is applied to either male or female facial data. Both the valence prediction results (0.3912 for male, 0.3714 for female and 0.3943 for gender-independent) and arousal prediction results (0.3711 for male, 0.3588 for female and 0.3878 for gender-independent) indicate that male facial behaviours are more correlated with their affect status. Moreover, it should be noted that the gender-independent experiment achieved better performance than gender dependent experiments. This might be caused by that fact that the gender-dependent experiments have less data for model training. In addition, these results might indicate that females and males do not display large variations when expressing their affect via facial behaviours in work-like settings, i.e., such small variations can not compensate the negative impact of the reduced number of training data on worker affect prediction models.

Gender	Male	Female	Both
Valence	0.3912	0.3741	**0.3943**
Arousal	0.3711	0.3588	**0.3878**

Table 5. The average UAR results achieved for different genders.

Feature representation and long-term modelling: We also specifically investigate the influence of different model configurations on periodical affect classification performance. As we can see from Table 6, the backbone that was pre-trained using naturalistic facial expression dataset achieved slightly better affect recognition UAR results than the backbone that was pre-trained using a facial AU dataset, both of which clearly are higher than the chance-level prediction. This means both facial expression and AU-related facial features obtained from the workers' faces are correlated with their self-reported affective state. Then, using facial features as the frame-level representation to construct clip-level facial behaviour representation is a superior way, as this setting achieved better UAR results for the recognition of both valence and arousal. We assume this is because frame-level features retain more affect-related facial cues than frame-level predictions, and thus during long-term modelling, frame-level feature-based clip-level representations can encode more affect-related temporal behavioural cues. Finally, simply choosing the mode of all frame-level predictions or the average feature of all frame-level features provided the worst results among all long-term modelling strategies, while spectral encoding achieved the best average performance for all baselines. This is because the encoded spectral representation contains multi-scale clip-level temporal dynamics.

	Strategy	Valence	Arousal
Backbone	GraphAU	0.3928	0.3770
	ResNet	**0.4064**	**0.3817**
Frame-level feature	Prediction	0.3866	0.3666
	Features	**0.4126**	**0.3921**
Long-term modelling	SE	0.4169	**0.3980**
	LSTM	**0.4182**	0.3785
	MODE/AVG	0.3635	0.3615

Table 6. The average results of different baseline configurations.

6 Conclusion

In this paper, we presented the first study that systematically investigated the face-based periodical valence and arousal analysis in work-like settings. More

specifically, this paper introduced a worker facial behaviour data acquisition protocol and the first cross-cultural human facial behaviour dataset in work-like settings. We also provided a set of deep learning-based baselines for face-based worker affect recognition. The results show that facial behaviours triggered by different tasks are informative for inferring valence and arousal states, but the performance is dependant on the task type and task setting. Our future work will focus on developing more advanced domain-specific loss functions and network architectures for multi-modal worker affect recognition.

Acknowledgement

This work is funded by the European Union's Horizon 2020 research and innovation programme project WorkingAge, under grant agreement No. 82623. S. Song and H. Gunes are also partially supported by the EPSRC/UKRI under grant ref. EP/R030782/1. Y. Luo contributed to this work while undertaking a summer research study at the Department of Computer Science and Technology, University of Cambridge.

References

1. Antoniadis, P., Pikoulis, I., Filntisis, P.P., Maragos, P.: An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3645–3651 (2021)
2. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018)
3. Borghini, G., Bandini, A., Orlandi, S., Di Flumeri, G., Aricò, P., Sciaraffa, N., Ronca, V., Bonelli, S., Ragosta, M., Tomasello, P., et al.: Stress assessment by combining neurophysiological signals and radio communications of air traffic controllers. In: International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 851–854. IEEE (2020)
4. Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry **25**(1), 49–59 (1994)
5. Du, Z., Wu, S., Huang, D., Li, W., Wang, Y.: Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition. IEEE Transactions on Affective Computing (2019)
6. Giorgi, A., Ronca, V., Vozzi, A., Sciaraffa, N., Di Florio, A., Tamborra, L., Simonetti, I., Aricò, P., Di Flumeri, G., Rossi, D., et al.: Wearable technologies for mental workload, stress, and emotional state assessment during working-like tasks: a comparison with laboratory technologies. Sensors **21**(7), 2332 (2021)
7. Guo, J., Lei, Z., Wan, J., Avots, E., Hajarolasvadi, N., Knyazev, B., Kuharenko, A., Junior, J.C.S.J., Baró, X., Demirel, H., et al.: Dominant and complementary emotion recognition from still images of faces. IEEE Access **6**, 26391–26403 (2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

9. Herreras, E.B.: Cognitive neuroscience; the biology of the mind. Cuadernos de Neuropsicología/Panamerican Journal of Neuropsychology **4**(1), 87–90 (2010)
10. Ilyas, C.M.A., Rehm, M., Nasrollahi, K., Madadi, Y., Moeslund, T.B., Seydi, V.: Deep transfer learning in human–robot interaction for cognitive and physical rehabilitation purposes. Pattern Analysis and Applications pp. 1–25 (2021)
11. Ilyas, C.M.A., Song, S., Gunes, H.: Inferring user facial affect in work-like settings. arXiv preprint arXiv:2111.11862 (2021)
12. Izard, C.E.: Human emotions. emotions, personality, and psychotherapy. New York: PlenumPress (1977)
13. Jenkins, J.M.: Self-monitoring and turnover: The impact of personality on intent to leave. Journal of Organizational Behavior **14**(1), 83–91 (1993)
14. Keltner, D.: Facial expressions of emotion and personality. In: Handbook of emotion, adult development, and aging, pp. 385–401. Elsevier (1996)
15. Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. International Journal of Computer Vision **127**(6), 907–929 (2019)
16. Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B.W., et al.: Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. IEEE transactions on pattern analysis and machine intelligence (2019)
17. Lerner, J.S., Li, Y., Valdesolo, P., Kassam, K.S.: Emotion and decision making. Annual review of psychology **66**, 799–823 (2015)
18. Lohse, M., Rothuis, R., Gallego-Pérez, J., Karreman, D.E., Evers, V.: Robot gestures make difficult tasks easier: the impact of gestures on perceived workload and task performance. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 1459–1466 (2014)
19. Luo, C., Song, S., Xie, W., Shen, L., Gunes, H.: Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. In: Proceedings of the Thirty-First International Conference on International Joint Conferences on Artificial Intelligence (2022)
20. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE transactions on affective computing **3**(1), 5–17 (2011)
21. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2017)
22. Mou, W., Gunes, H., Patras, I.: Alone versus in-a-group: A multi-modal framework for automatic affect recognition. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **15**(2), 1–23 (2019)
23. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). pp. 1–8. IEEE (2013)
24. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE transactions on pattern analysis and machine intelligence **37**(6), 1113–1133 (2014)
25. Scherer, K.R.: What are emotions? and how can they be measured? Social science information **44**(4), 695–729 (2005)

26. Song, S., Jaiswal, S., Shen, L., Valstar, M.: Spectral representation of behaviour primitives for depression analysis. IEEE Transactions on Affective Computing (2020)
27. Song, S., Shen, L., Valstar, M.: Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In: 2018 13th IEEE FG (2018). pp. 158–165. IEEE (2018)
28. Song, S., Sánchez-Lozano, E., Kumar Tellamekala, M., Shen, L., Johnston, A., Valstar, M.: Dynamic facial models for video-based dimensional affect estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
29. Tsai, Y.F., Viirre, E., Strychacz, C., Chase, B., Jung, T.P.: Task performance and eye activity: predicting behavior relating to cognitive workload. Aviation, space, and environmental medicine **78**(5), B176–B185 (2007)
30. Yang, C., Xu, Y., Shi, J., Dai, B., Zhou, B.: Temporal pyramid network for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 591–600 (2020)
31. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32**(10), 692–706 (2014)
32. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image and Vision Computing **29**(9), 607–619 (2011)
33. Zimmerman, R.D.: Understanding the impact of personality traits on individuals'turnover decisions: A meta-analytic path model. Personnel Psychology **61**(2), 309–348 (2008)