

Received March 23, 2022, accepted April 3, 2022, date of publication April 7, 2022, date of current version April 20, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3165622

Privacy Aware Affective State Recognition From Visual Data

M. SAMI ZITOUNI^{1,2}, (Member, IEEE), PETER LEE³, UICHIN LEE³,
LEONTIOS J. HADJILEONTIADIS^{1,2,4}, (Senior Member, IEEE),
AND AHSAN KHANDOKER^{1,2}, (Senior Member, IEEE)

¹Department of Biomedical Engineering, Khalifa University, Abu Dhabi, United Arab Emirates

²Health Engineering Innovation Center, Khalifa University, Abu Dhabi, United Arab Emirates

³Graduate School of Knowledge Service Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

⁴Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

Corresponding author: M. Sami Zitouni (mohammad.zitouni@ku.ac.ae)

ABSTRACT Affective state recognition is a key component of any system equipped with emotional awareness and intelligence. The ability of recognizing emotions allows the machine to better understand the user requirements, and guide its decision and response, thus establishing a more connected relationship with the human. It is usually assumed that emotion recognition is mainly determined by face features from visual data, which impose the dilemma of invading the privacy of the user and capturing their identity, which is unacceptable by many people, especially in public human-machine interaction (HMI) setups. On the other hand, bodily reactions and background context can provide enough emotional clues visually and are less susceptible to contextual influences compared to facial expression. Consequently, this paper investigate the recognition of affective state from visual data captured during a naturalistic conversation with similar perspective to HMI. The faces were masked to conceal the identity of the users. A deep learning recognition model based on a combined Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) architecture is employed to classify the user's affective state into two levels of arousal and valence, as well as their quadrant combinations. The experiments were conducted using two different labeling schemes mimicking the self and conversation partner perspectives. The results shows that affective state recognition from masked data using the proposed model can achieve comparable performances (up to 96.82%, 95.91%, and 91.52% for arousal, valence, and quad classes recognition, respectively) in comparison to the use of raw data with facial expressions. This paves the way for privacy aware emotion recognition systems that could be widely accepted by the users.

INDEX TERMS Affective computing, arousal-valence, dimensional emotions, visual signals, privacy preserving, CNN-LSTM.

I. INTRODUCTION

Affective state recognition plays a vital role in natural human-machine interaction (HMI). The accurate recognition can provide in-depth user experiences, such as smart home, education, health monitoring, and so on. Among various approaches to recognize human emotions, facial expressions are often adopted and analyzed through deep learning algorithms [15]–[17]. However, faces convey characteristics of the person, such as age, gender, and identity, which pose threats to user privacy. As recent advancements in facial recognition technology provide ways to screen

persons and even a form of identification, there is a growing demand for de-identification of biometric data to protect user privacy [18].

Another source of affective information that does not convey user identity can be upper body movements of individuals and the background context. As a means of human communication, body posture, gestures, hand, and head movement also convey a significant amount of information [19], [20]. For instance, people tend to stay put their arms on the table for neutral emotion, but extend or move arms closer to their faces when they feel happy, sad, or fear etc. Some body movements are subjective as expression of emotion varies with people and cultural bias can interfere [21]; however, there is general consensus on intentions under body languages. In addition, for


The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang .

TABLE 1. Summary of related works.

Work	Classifier	Privacy preserving	Dataset	Results	Advantages	Limitations
Pentyala <i>et al.</i> , 2021 [1]	CNN	Secure multi-party computation protocols	RAVDESS [2]	56.8% for 7 classes	End-to-end privacy-preserving classification pipeline	Single frame based and no visual masking
Narula <i>et al.</i> , 2020 [3]	CNN	Adversarial learning approach, which suppresses identity-specific information	JAFFE [4], YALE [5], IEMOCAP [6]	57.52-92.62%	Preserving emotion-specific information while reducing privacy-sensitive information	Image-based, unnaturalistic emotion identification
Petrova <i>et al.</i> , 2020 [7]	CNN	Group-level based emotion recognition	VGAF [8]	59.13% for 3 classes	No individual-based feature as input	Image-based, visual identity of individuals are not masked
Hossain <i>et al.</i> , 2019 [9]	CNN, SVM	Secret sharing scheme	RML [10], eNTER-FACE [11]	82.3-87.6%	Image and speech signals are used, ensuring user's privacy when sharing data	The facial and speech data used include user identity
Jiang <i>et al.</i> , 2017 [12]	K-NN	Face scrambling	JAFFE [4], MUG [13], CK+ [14]	40.71-95.24%	Allows privacy-protected facial expression recognition	Image-based with handcrafted features, unnaturalistic setup

some cases, body posture overrides that of facial expression in terms of emotion recognition as it better conveys contextual situations [22], [23]. Further, visual or background context, which is the background and surrounding influences from many modalities, such as scene gist information, faces of surrounding people, and perceiver personality traits, is essential information in the emotional experience. Aside from a person's facial expressions and body information, surrounding context contributes directly to the perception of emotion, and thus contains cues to infer affective information over time, especially when temporal information is captured [24], [25].

There are two major perspectives on emotion recognition; 1) discrete emotion that relies on emotive language by the Ekman model [26], and 2) dimensional emotion that differentiate in terms of arousal, valence, and dominance. However, discrete emotions often face difficulties in categorizing with different languages, for instance, it is hard to find specific matches in other languages [27]. In the case of dimensional approach, different emotions can be represented through spanning arousal, valence, and dominance space. Arousal indicates whether activated or not and the degree of emotion; valence determines the positive or negative feelings; and dominance specifies the control or not [28]. Further, the representation of emotion does not require specific categorization which may account for different emotions from individuals [29]. Lastly, arousal and valence are mainly adopted for continuous annotation and labeling of the emotion as these two dimensions explains most of the variability.

In this paper, we propose a framework for affective state classification in the arousal and valence space, using body gestures and background context captured from video recordings with completely blurred faces. This study is based on visual recordings during a naturalistic controversial debate between pairs, where each partner's emotions were annotated frequently from different perspectives. A neural network classification model based on a combined Convolutional Neural Network and Long Short-Term Memory (CNN-LSTM) architecture is adapted for accurate and real time classification of the emotions into two levels of arousal and valence,

in addition to their quadrant combinations. Further, the raw video data without face masking were processed to compare the robustness of emotion recognition with body movements and the annotation from self and partner's were compared. We argue that masking the identity defining aspect, which is the face in this case, does not degrade the affective state recognition performance as can be expected, since upper body gestures and background context captured via videos can be a great source of emotional cues.

A. RELATED WORK

Pentyala *et al.* [1] proposed an implementation of single frame method for privacy preserving CNN based emotion classification in videos. It allowed a party to infer a label from a video without requiring the owner to disclose his/her video unencrypted, as well as not requiring the reveal of the classifier parameters to user. This is done through applying private image classification protocols for single frame selection and label aggregation. In this method, the identity of the user is still visually revealed through the recorded video, making it unsuitable for public HMI applications. Additionally, being a single frame-based method, eliminates the temporal factor of the data, which can contain a great amount of cues for emotions. Narula *et al.* [3] presented a framework that recognizes emotions through user anonymization. In this approach, emotion-specific information was preserved, while user-dependent convolutional kernel of CNN was eliminated, thus reducing user re-identification. This work is image-based and was tested on datasets with unnaturally distinctive emotional facial expressions. Petrova *et al.* [7] proposed a group emotion detection non-individual approach that is privacy-safe. This method was based on frugal modeling, where only global image cues were processed, in addition to mixing available datasets with a synthetic one. This work is also based on still images, ignoring the temporal aspect, and the input images include individuals' identity visual data, although processed within groups. Hossain and Muhammad [9] proposed emotion recognition system based on edge-cloud, where Internet of Things (IoT) devices capture visual

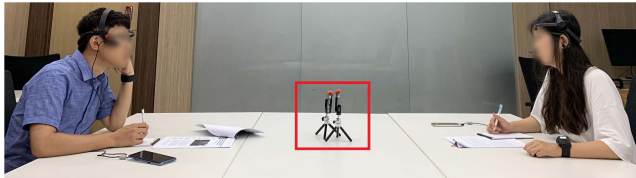


FIGURE 1. Data collection setup of K-EmoCon dataset [30] with capturing cameras mounted in 2nd-person point of view.

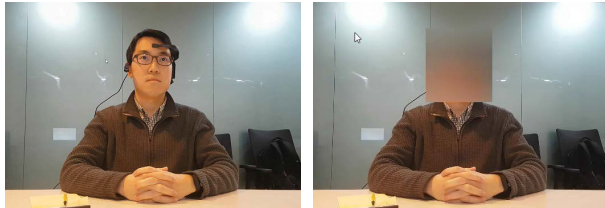


FIGURE 2. Example raw frame and its masked version from K-EmoCon dataset [30] of a participant who consented for disclosure of data.

and speech signals from a user and use secret sharing scheme for distribution to different edge clouds, to ensure privacy. A CNN model was used to extract features from image and speech signals, and an SVM was used for classification. This work ensures private data sharing, while the identity of the user is not covered in neither image (masking) nor speech (distortion) signals. Jiang *et al.* [12] introduced many graph embedding technique to identify discriminative patterns from the subspaces of chaotic patterns, in order to recognize emotions in images with scrambled facial expression, using nearest neighbor classifier (K-NN). In terms of preserving privacy, this approach used facial scrambling which covers user's identity. Nonetheless, this work is image-based that used handcrafted featured, tested on still unnaturalistic identifiable facial expressions.

II. METHODOLOGY

A. DATASET

This framework is aimed towards the recognition of humans affective state from their visual appearance, during their communication and interaction with other humans or machines as in HMI applications. Thus, K-EmoCon [30] dataset is adopted in this work, which is a publicly available multi-modal resource of affective information. It consists of videos, which are used in this work, as well as audio recordings and various physiological signals captured with multiple wearable sensors. The data were collected and annotated during a naturalistic conversation in form of turn-taking 10-minute debates on a controversial social issue in environments with controlled temperature and illumination. Facial expressions and movements in the upper body were captured from the 2nd-person point of view, which is also the natural perspective of a machine in HMI, using two smartphones mounted on tripods in the middle of the table facing each participant, as shown in Fig. 1. 21 participants {P2, P3, P4, P5, P7, P8, P9, P10, P13, P15, P19, P20, P21, P22, P23, P24, P25, P26, P29, P30, P31} out of 32 provided

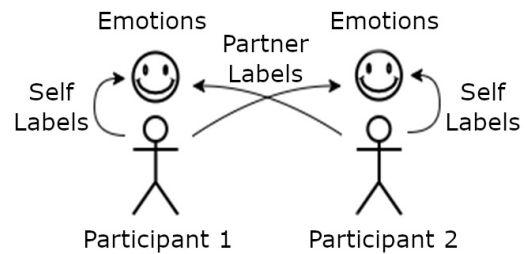


FIGURE 3. Illustration of self and partner annotations during a conversation between two participants.

consent (65% of participants) for the disclosure of data with personally identifiable information, which are the videos used in this work. This, for example, confirms people's concern or acceptance of allowing capturing their identity even in data collection setups. Thus, a total of 223.35 minutes of footage are available, with resolution of 1920×1080 and frame rate of 30 fps.

B. FACE MASKING

Privacy is a critical concern in visual based recognition systems, as most people refrain from interacting with machines that collect their personally identifiable information. Thus, we investigate the visual recognition of human affective state mainly from their body gestures without disclosing the most identifiable feature, which is the face. This is done by masking the face expressions and appearance before collecting and processing the visual data for a personalized HMI experience. In this study, all the publicly available 21 videos in the K-EmoCon [30] dataset were masked priority. In such a setup, the masking can be done either by simply specifying a certain region in the frame where the face of the participant is located and dilate then mask it by applying a high intensity blurring filter to the specified region, which works here since the capturing camera and the pose of the participants are fixed, thus no extra processing is needed on the videos. The other way is to use a face detector to obtain the face region, and then applying a blurring filter on dilated region to mask the face in the current frame. To avoid revealing the participant's face in frames where the detector misses, the last successfully detected face region will be maintained masked. Fig. 2 shows example raw frame as well as its masked version using a heavy blurring filter, which are used in this framework.

C. AFFECTIVE STATE ANNOTATION

Emotions of the participants were annotated during the debate period every 5 seconds from different perspectives. Here, self (the participants rating themselves) and partner (the debate partners rating each other) annotations are adopted, as well as combined annotation combining both ratings. Fig. 3 illustrates the annotation scheme [31]. The emotions were annotated based on arousal and valence affective dimensional emotional model as in Russell's circumplex model of affect [32], and they were measured with

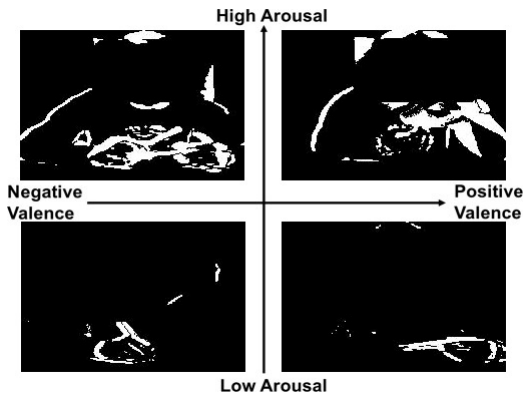


FIGURE 4. Example of foreground pixels from masked recording frames of P5 while expressing different affective states on the arousal-valence space.

a Likert scale from 1 to 5. Accordingly, the emotions are classified by the model based on the level of arousal and valence into high (H) and low (L). Therefore, the self and partner annotations in Likert scale ratings are converted into H and L according to a mid-value of 2.5 (L: 1-2, H: 3-5). Moreover, the emotions are categorized into one of quad classes combining the arousal and valence levels, which are and high arousal high valence (HAHV), high arousal low valence (HALV), low arousal high valence (LAHV), low arousal low valence (LALV). For the combined annotations, the self and partner ratings are accumulated and re-scaled into 1 to 9, then converted into H and L based on a mid-value of 4.5 (L: 1-4, H: 5-9). Fig. 4 displays detected foreground from instances of participant's P5 recording, laid on each quad of the arousal-valence space according to the partner's ratings.

D. RECOGNITION MODEL

The recognition of the affective state from visual data in this framework is performed using a combined CNN-LSTM architecture. A block diagram that illustrates the proposed recognition model is shown in Fig. 5. First, the input video is divided into batches of size ℓ frames. In case of 5s classification, which is the annotation interval in the K-Emocon dataset, ℓ will be 150 frames ($5s \times 30fps$). The input batches are cropped according to the region where the human is located, and resized to match the input size of the CNN network. A sequence input layer is used to take the image sequence (video batch).

Sequence features extraction layers convert the input video batch into sequences of feature vectors based on a pre-trained CNN. First, a sequence folding layer is used to get an array of images out of the video patch, allowing the convolutional operations to be applied on each frame independently. The pre-trained convolutional layers of a CNN network are used for feature extraction by getting the activations of each frame. In this study, the convolutional layers of GoogleNet network are used. The output video features sequence is connected to sequence unfolding layer, which reestablishes the input sequence structure, and flatten layer, which converts

the pooled feature maps into 1D vectors. Subsequently, the output will be 1024 feature vectors (corresponding to last pooling layer of the CNN) of size ℓ .

For time dependent sequence feature classification of the input video batch, LSTM layers are utilized. The LSTM network scheme used in this model consists of three bi-directional layers, each followed by a dropout layer. The first Bi-LSTM layer has 1000 hidden units while the second and third have 500 hidden units. The first two Bi-LSTM layers return sequence, and are followed by dropout layers with 0.8 probability. The last Bi-LSTM layer returns state (last time stamp) and is followed by a dropout layer with 0.2 probability. Finally, a fully connected layer followed by a softmax layer are used for classification result. This model was trained for three independent tasks, arousal classification, valence classification, and quad arousal-valence classification. Thus, the output will have either two states, corresponding to the two levels of arousal and valence, or four states when used for quad recognition, where the output classes are as was described in Section II-C.

E. IMPLEMENTATION

The training and testing of the presented framework was performed in Matlab 2021a. In the training phase, the following settings were used. GoogLeNet convolutional layers were loaded and the parameter were fixed without performing fine-tuning. The training options had the initial learning rate set to 0.0005, the minimum batch size to 32, and the gradient threshold was 2. The recognition network was trained for each experiment with 35 epochs, while shuffling the data every epoch. These hyper-parameters were selected after preliminary experimentation and testing of the network.

III. RESULTS & DISCUSSION

To validate the proposed framework, experiments were conducted in a 4-fold cross-validation setup, where the same validation recipe was followed in all tasks to ensure a fair comparison. The proposed recognition model was trained and tested using raw videos as well as masked videos as inputs for personally identifiable information preservation to compare the performance of model when used in privacy aware systems. The same implementation setup and hyper-parameters were used in all tasks. Additionally, two annotation perspectives were considered in training and validation, which are self and partner, to investigate whether there is a direct reflection of the perspective of the emotion annotation on the results in visual data based emotion recognition.

Table 2 shows the accuracy of the proposed affective state recognition framework using different testing setups. The recognition model was trained and tested for arousal, valence, and quad states recognition using self as well as partner annotations, for both raw and masked videos. Furthermore, the training and testing were performed with various sizes of input video batches (different ℓ values) to study the effect of using longer video batches from the same dataset, where the annotations were performed each 5s. Thus, for periods more

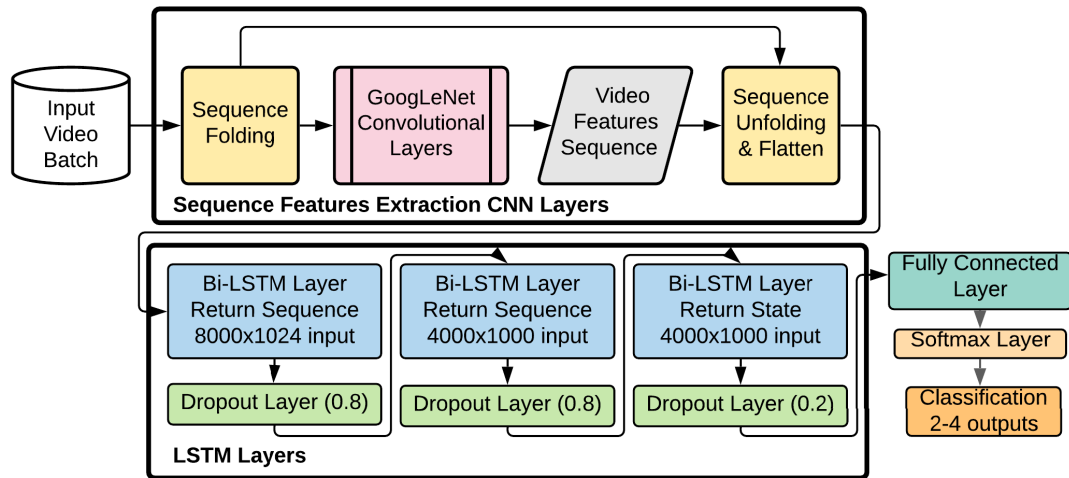


FIGURE 5. Illustration of the proposed visual based affective state recognition model.

TABLE 2. Emotion recognition accuracy (%) results of the proposed model using raw and masked input videos, with different annotation perspectives, and input window lengths (number of frames ℓ in each batch).

		Self Annotation			Partner Annotation		
		Arousal	Valence	Quad	Arousal	Valence	Quad
5s ($\ell = 150$)	Raw	88.79	90.3	81.82	93.94	94.24	89.24
	Masked	90.45	89.24	82.27	95.15	93.48	89.09
15s ($\ell = 450$)	Raw	89.24	90.45	83.79	95	96.21	91.36
	Masked	91.67	90.61	83.33	96.21	95.76	90.76
25s ($\ell = 750$)	Raw	89.09	92.73	85.76	95.15	97.12	92.42
	Masked	91.36	91.67	84.39	96.21	95.91	91.52
35s ($\ell = 1050$)	Raw	92.12	91.97	85.45	95.61	97.73	93.64
	Masked	92.58	90	83.18	96.82	95.61	90.45
45s ($\ell = 1350$)	Raw	90.61	92.73	85.61	93.48	97.58	93.64
	Masked	91.06	90.91	82.12	95.3	93.94	90.61

than 5s, the rating values were accumulated and re-scaled to identify the affective state category (arousal-valence level) accordingly. Additionally, to maintain similar number of samples in all tests, overlapping in video batches was present in longer windows, where an increment of 5s remained between the starting points of video batches. This was investigated as, in K-EmoCon dataset, we calculated the average period of time that an affective state level remains unchanged for all participants, and it was found to be 135s (190s for H and 80s for L) in arousal, and 110s (180s for H and 40s for L) in valence.

First, looking at the emotion recognition results with 5s window length, it can be observed that when using the partner annotation, the accuracy is much higher in comparison to the self annotation in all tasks. For both raw and masked input videos, the arousal accuracy is increased by around 5%, the valence by 4%, and the Quad classes by 7%. The same observation can be made for longer input window lengths. This can be inferred by the fact that from partner perspective, the affective state rating is performed mainly based on the visual expressions and body language, which is the perspective and type of information used in the proposed framework, thus making the partner annotation

the more appropriate ground-truth. On the other hand, looking at the results using different input batch sizes, it can be noticed that at longer windows the accuracy increases, then decreases in some cases at 45s like arousal with partner annotations. The highest accuracy of arousal recognition tests were obtained with 35s as well as for valence and quad recognition from raw data with partner annotation. Using 25s input window, the best results of valence and quad recognition with self annotations were obtained, and for valence and quad recognition from masked data with partner annotation. Thus, it can be deduced from these results that in such scenarios as the debate conversation of the used dataset, 25s to 35s is the most reasonable time window at which changes in the affective state of the participants can be captured, while 5s window is too short for frequent change in the participants emotional state, and fluctuation in the annotations can be a result of human error. On the contrary, in a period of more that 45s, some changes in the participants state can be missed or flattened, in case more than one actual change in the emotion occurred within this longer period.

The illustrated results in Table 2 prove that the proposed affective stated recognition framework can be employed in privacy aware system with data masking the personally

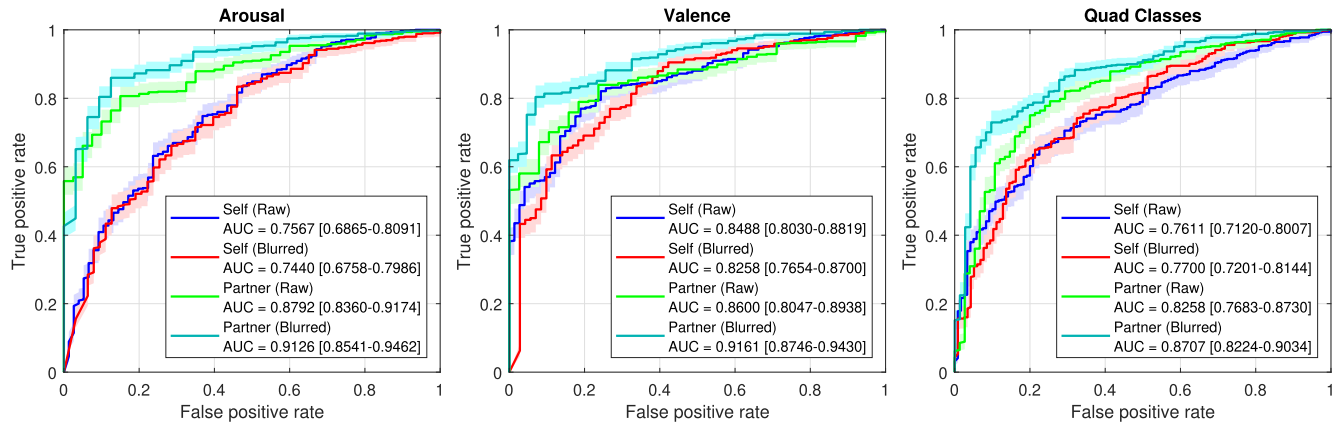


FIGURE 6. ROC curves of affective state recognition results for different annotations, and AUC measured with 95% confidence interval.

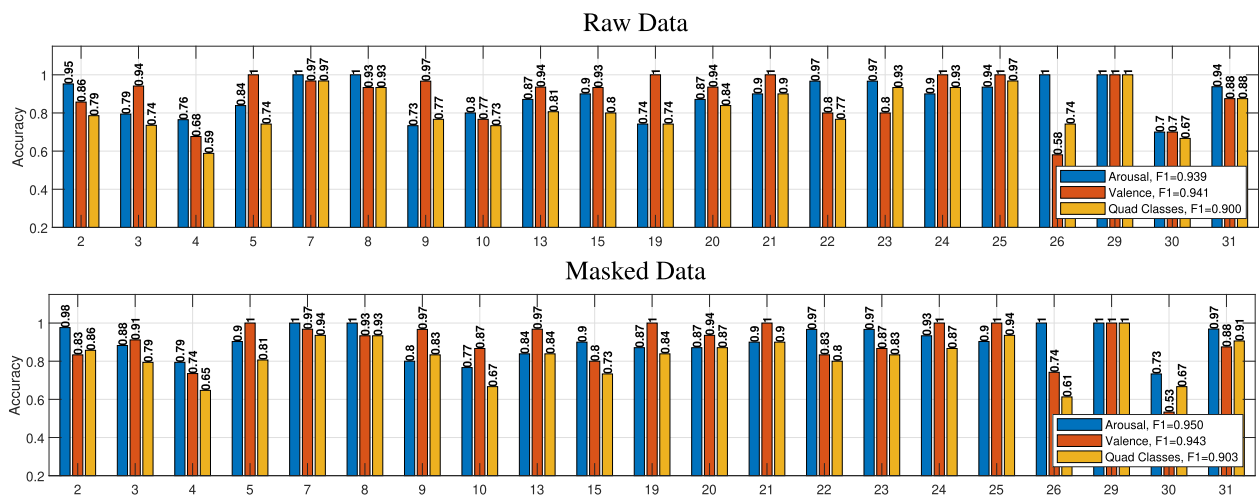


FIGURE 7. Recognition accuracy for each participant and overall F1 score with self annotation.

identifiable information of users. When used with masked input data, the obtained results are very close and competitive to the case of raw input data with visible facial expressions. Using masked data at the base annotation period, the arousal recognition accuracy is around 1.5% higher, valance is 1% lower for both annotations, and Quad is almost the same. The slight improvement in arousal recognition can be due to the fact that the arousal state is more expressed by the body language in comparison to the facial expressions. The results indicate that the emission of the facial expression information allowed more clear distinction of the arousal state for some participants, as will be seen later in participant wise accuracies. This demonstrates that with the absence of facial expressions, the temporal element of the model that captures the changes in body posture, can be more reliable in inferring the arousal state. For valence, the effect was vise versa, but still very minimal, which shows that in applications mimicking the perspective of the data capturing, and in natural conversation sittings, the proposed framework can perform well, while preserving the personally identifiable information

of the interacting human. The best results obtained with masked data for arousal, valence, and quad states recognition are 96.82%, 95.91%, and 91.52% respectively, in comparison to 95.61%, 97.73%, and 93.64% with raw data.

Fig. 6 displays ROC curves of the proposed affective state recognition framework when using raw data against masked data with self and partner annotations. Moreover, the AUC values are calculated with 95% confidence intervals highlighted. Interestingly, the highest AUC values obtained for all arousal, valence, and quad recognition were using masked (blurred) data with partner annotations (0.9126, 0.9161, and 0.8707 respectively). This demonstrates the capability of the proposed framework to recognize the affective state from visual data, while preserving the personally identifiable information. Additionally, as it was observed previously, partner annotation with visual data leads to a higher performance, especially here in the arousal recognition case.

The recognition model performance on each participant's raw and masked data is shown in Fig. 7 and Fig. 8 for self

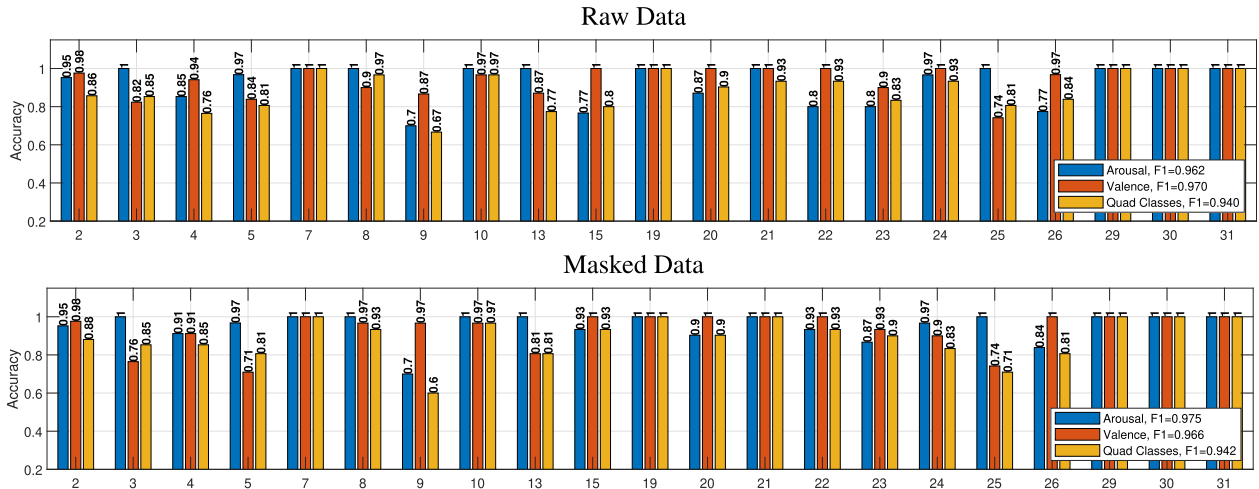


FIGURE 8. Recognition accuracy for each participant and overall F1 score with partner annotation.

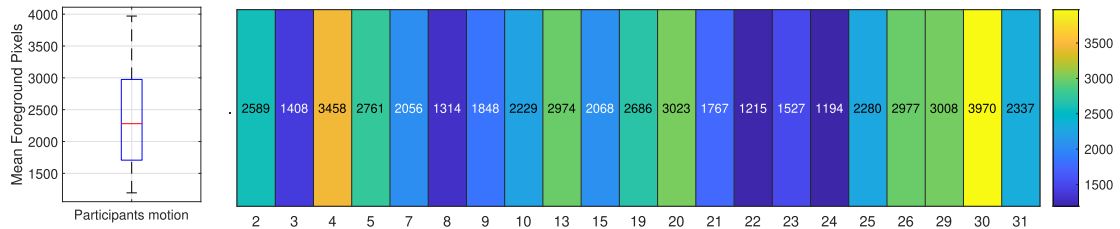


FIGURE 9. A box plot showing the distribution of mean foreground pixels for participants' videos during the recording session, quantifying their body gestures, with the corresponding heatmap.

and partner annotations. Additionally, the overall arousal, valence, and quad classification F1 score is estimated for each scenario. The improvement of the results from using the partner perspective can clearly be noticed from the results of participants such as *P4*, *P10* and *P30*, as well as the higher F1 scores. Considering Fig. 8, the same aforementioned conclusion can be drawn, where F1 score with masked data is slightly higher for arousal recognition (0.975 against 0.962), and slightly lower for valence recognition (0.966 against 0.970). This pattern can be clearly seen in the arousal results of *P4*, *P15*, *P22* and *P26*, as well as the valence results of *P3*, *P5*, *P13* and *P24*. However, this is not necessarily the case for other participants. Overall, the obtained results show that the system is capable of accurate recognition whether the facial expressions are masked or not.

Finally, to investigate the body movements that allowed the private recognition of the affective state during the recording sessions, the motion of each participant was measured through foreground detecting using adaptive Gaussian mixture models [33]. The foreground mask was extracted for each frame in the input video patches. Then movement was quantified by the number of pixels in the foreground masks. Fig. 9 displays the quantification of motion using the mean number of foreground pixels during each participant's recording session, in form of a box plot of the distribution across all participants, and a heatmap showing each participants mean

foreground pixels count against others. Using the information obtained from the box plot and heatmap, the participants can be categorized according to their average movement using thresholds around the upper and lower ends of the box (2700 and 1800 pixels respectively). Thus, they are divided into ones who exhibited low movement (*P3*, *P8*, *P21*, *P22*, *P23*, *P24*), medium movement (*P2*, *P5*, *P7*, *P9*, *P10*, *P15*, *P19*, *P25*, *P31*), and high movement (*P4*, *P13*, *P20*, *P26*, *P29*, *P30*). According to the results in the masked data plot of Fig. 8, the average arousal and valence recognition accuracies for low movement participants are 96.2% and 92.7%, while for medium movement participants they are 95% and 93%, and for high movement participants they are 94.2% and 95.3%, accordingly. It is observed that the average arousal recognition accuracy decreases across participants with higher amount of movement, while the average valence accuracy increases.

IV. CONCLUSION

In this work, privacy aware affective state recognition from visual data was investigated. Visual cues of body movements and background context were captured from videos with masked people that preserve personally identifiable information, and a comparative study was performed against the use of raw videos with facial expressions. A combined CNN-LSTM network was proposed for a robust visual based

affective state recognition according to the arousal-valence space. The framework was verified on data collected during a naturalistic conversation, and the comparative analysis was performed using two different annotation perspectives. The results showed that the performance of the model when trained using partner annotations with visual data, was higher than when used with self annotations, as the conversation partner depends mainly on the visual clues to annotate, which is not the case when people annotate themselves. Further, the affective state recognition using masked data achieved competitive and in some cases superior results, especially in arousal recognition, in comparison to the use of raw footage with facial expressions for the same tasks. This is due to the fact that adequate emotional cues are embodied in body gestures and background context. Thus, this work shows the potential of a visual based emotion recognition system that is more considerate towards users' privacy concerns, and can be widely acceptable in HMI and other applications.

REFERENCES

- [1] S. Pentyala, R. Dowsley, and M. De Cock, "Privacy-preserving video classification with convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8487–8499.
- [2] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.
- [3] V. Narula, K. Feng, and T. Chaspari, "Preserving privacy in image-based emotion recognition through user anonymization," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 452–460.
- [4] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, "The Japanese female facial expression (JAFFE) database," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 14–16.
- [5] *Yale Face Database*. Accessed: Mar. 22, 2022. [Online]. Available: <http://vision.ucsd.edu/content/yale-face-database>
- [6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
- [7] A. Petrova, D. Vaufreydz, and P. Dessus, "Group-level emotion recognition using a unimodal privacy-safe non-individual approach," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 813–820.
- [8] G. Sharma, S. Ghosh, and A. Dhall, "Automatic group level affect and cohesion prediction in videos," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2019, pp. 161–167.
- [9] M. S. Hossain and G. Muhammad, "Emotion recognition using secure edge and cloud computing," *Inf. Sci.*, vol. 504, pp. 589–601, Dec. 2019.
- [10] Y. Wang and L. Guan, "Recognizing human emotional state from audio-visual signals," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 936–946, Aug. 2008.
- [11] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2006, p. 8.
- [12] R. Jiang, A. T. S. Ho, I. Cheheb, N. Al-Maadeed, S. Al-Maadeed, and A. Bouridane, "Emotion recognition from scrambled facial images via many graph embedding," *Pattern Recognit.*, vol. 67, pp. 245–251, Jul. 2017.
- [13] N. Aifanti and A. Delopoulos, "Linear subspaces for facial expression recognition," *Signal Process., Image Commun.*, vol. 29, no. 1, pp. 177–188, Jan. 2014.
- [14] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (Workshops)*, Jun. 2010, pp. 94–101.
- [15] I. M. Revina and W. R. S. Emmanuel, "A survey on human face expression recognition techniques," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 33, no. 6, pp. 619–628, Sep. 2018.
- [16] S. Wu, Z. Du, W. Li, D. Huang, and Y. Wang, "Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 40–48.
- [17] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, Jan./Mar. 2016.
- [18] M. Shopon, S. N. Tumpa, Y. Bhatia, K. N. P. Kumar, and M. L. Gavrilova, "Biometric systems de-identification: Current advancements and future directions," *J. Cybersecur. Privacy*, vol. 1, no. 3, pp. 470–495, Aug. 2021.
- [19] K. Lang, M. M. Dapelo, M. Khondoker, R. Morris, S. Surguladze, J. Treasure, and K. Tchanturia, "Exploring emotion recognition in adults and adolescents with anorexia nervosa using a body motion paradigm," *Eur. Eating Disorders Rev.*, vol. 23, no. 4, pp. 262–268, Jul. 2015.
- [20] B. de Gelder, A. W. de Borst, and R. Watson, "The perception of emotion in body expressions," *Wiley Interdiscipl. Rev., Cognit. Sci.*, vol. 6, no. 2, pp. 149–158, Mar. 2015.
- [21] K. Lander and N. L. Butcher, "Recognizing genuine from posed facial expressions: Exploring the role of dynamic information and face familiarity," *Frontiers Psychol.*, vol. 11, p. 1378, Jul. 2020.
- [22] M. Kayyal, S. Widen, and J. A. Russell, "Context is more powerful than we think: Contextual cues override facial cues even for valence," *Emotion*, vol. 15, no. 3, p. 287, 2015.
- [23] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 338, pp. 1225–1229, 2012.
- [24] Z. Chen and D. Whitney, "Tracking the affective state of unseen persons," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 15, pp. 7559–7564, Apr. 2019.
- [25] Y. Huang, H. Wen, L. Qing, R. Jin, and L. Xiao, "Emotion recognition based on body and context fusion in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3609–3617.
- [26] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Sci. Inf.*, vol. 44, no. 4, pp. 695–729, 2005.
- [27] J. A. Russell, "Culture and the categorization of emotions," *Psychol. Bull.*, vol. 110, no. 3, p. 426, 1991.
- [28] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, no. 3, pp. 273–294, 1977.
- [29] A. Anderson, T. Hsiao, and V. Mettsis, "Classification of emotional arousal during multimedia exposure," in *Proc. 10th Int. Conf. Pervasive Technol. Rel. Assistive Environments*, Jun. 2017, pp. 181–184.
- [30] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Sci. Data*, vol. 7, no. 1, p. 293, Dec. 2020.
- [31] M. S. Zitouni, C. Y. Park, U. Lee, L. Hadjileontiadis, and A. Khandoker, "Arousal-valence classification from peripheral physiological signals using long short-term memory networks," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 686–689.
- [32] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.
- [33] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*. Boston, MA, USA: Springer, 2002, pp. 135–144.



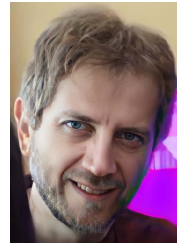
M. SAMI ZITOUNI (Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical and computer engineering from Khalifa University (KU), Abu Dhabi, United Arab Emirates, in 2015 and 2019, respectively. He conducted his studies with the KU Center for Autonomous Robotic Systems (KUCARS) and the Visual Signal Analysis and Processing Center (VSAP). He is currently a Postdoctoral Fellow in biomedical engineering at KU as part of KU—Korean Advanced Institute of Science and Technology (KU-KAIST) Joint Research Center. His research interests include artificial intelligence, machine learning applications, affective computing, computer vision, signal processing, and embedded systems.



PETER LEE received the B.S.E. and M.S.E. degrees in bioengineering from the University of Pennsylvania, Philadelphia, PA, USA, in 2013, and the Ph.D. degree in bio and brain engineering from KAIST, Daejeon, South Korea, in 2019. He is currently a Contract Researcher at the KAIST Institute for Health Science and Technology, KAIST. His research interests include neuroimaging, digital health-care, wearable device, artificial intelligence, and affective computing.



UICHIN LEE received the B.S. degree in computer engineering from Chonbuk National University, in 2001, the M.S. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), in 2003, and the Ph.D. degree in computer science from UCLA, in 2008. He continued his studies at UCLA as a Postdoctoral Research Scientist (2008–2009) and then worked for Alcatel-Lucent Bell Labs as a Member of Technical Staff (till 2010). He is currently an Associate Professor with the School of Computing, KAIST. His research interests include human–computer interaction (HCI), social computing, and ubiquitous computing.



LEONTIOS J. HADJILEONTIADIS (Senior Member, IEEE) received the Diploma degree in electrical engineering and the Ph.D. degree in electrical and computer engineering (ECE) from the Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece, in 1989 and 1997, respectively, the Ph.D. degree in music composition from the University of York, York, U.K., in 2004, and the Diploma degree in musicology from AUTH, in 2011. He is with ECE-AUTH (Professor) and the Department of Biomedical Engineering (Professor/Chair), Khalifa University, Abu Dhabi, United Arab Emirates. His research interests include advanced signal processing, machine learning, biomedical engineering, affective computing, and active and healthy ageing.



AHSAN KHANDOKER (Senior Member, IEEE) received the Ph.D. degree in electronics and biomedical engineering from the Muroran Institute of Technology, Japan, in 2004, followed by an Australian Research Council Fellowship with the Department of Electrical and Electronic Engineering, The University of Melbourne, Australia. He is currently an Associate Professor of biomedical engineering at Khalifa University, Abu Dhabi, United Arab Emirates. His research projects are funded by the Abu Dhabi Department of Education and Knowledge, the Bill and Melinda Gates Foundation, the Australian Research Council, and the Khalifa University internal funds in cardiac and mental health monitoring research area in collaboration with Cleveland Clinic Abu Dhabi and several key international medical research facilities in Australia, Germany, and Japan. His research interests include sleep, diabetes, fetal medicine, psychiatry, biomechanics, and bioinstrumentation; and bio-signal processing and circuits and nonlinear modeling.

...