A Deep Learning Approach Towards Multimodal Stress Detection

Cristian Paul Bara, Michalis Papakostas, and Rada Mihalcea

University of Michigan, Electrical Engineering & Computer Science
Ann Arbor MI 48109, USA
{cpbara,mpapakos,mihalcea}@umich.com

Abstract. Several studies that emerged from the fields of psychology and medical sciences during recent years have highlighted the impact that stress can have on human health and behavior. Wearable technologies and sensor-based monitoring have shown promising results towards assessing, monitoring and potentially preventing high-risk situations that may occur as a result of fatigue, poor health, or other similar conditions caused by excessive amounts of stress. In this paper, we present our initial steps in developing a deep-learning based approach that can assist with the task of multimodal stress detection. Our results indicate the promise of this direction, and point to the need for further investigations to better understand the role that deep-learning approaches can play in developing generalizable architectures for multimodal affective computing. For our experiments we use the MuSE dataset – a rich resource designed to understand the correlations between stress and emotion – and evaluate our methods on eight different information signals captured from 28 individuals.

Keywords: multimodal stress detection · representation learning · affective computing · deep learning

1 Introduction

Stress is a normal reaction of the body, mostly observed under situations where we struggle to cope with the conditions or the changes that occur in our environment [8]. Its effects and symptoms affect our body both physically as well as mentally and emotionally. Stress is thus playing a very significant role towards shaping our overall behavior, well-being and potentially our personal and professional success [16].

Affective computing is the field of science that studies and develops technologies able to capture, characterize and reproduce affects, i.e., experiences of feelings or emotions. It is a highly interdisciplinary domain, mostly influenced by the fields of computer science, psychology and cognitive science and was initially introduced in the late 90s by Rosalind Piccard [20]. Contributions of affective computing have traditionally played a very important role towards designing more engaging and effective Human-Computer-Interaction systems that can adapt their responses according to the underlying human behavior [17, 22].

 In: N. Chhaya, K. Jaidka, J. Healey, L. H. Ungar, A. Sinha (eds.): Proceedings of the 3rd Workshop of Affective Content Analysis, New York, USA, 07-FEB-2020, published at http://ceur-ws.org

However, despite the dramatic evolution of affective computing and computer science in general during the last decade, capturing and analysing stress factors and their effects on the body remains a very challenging problem primarily due to the multidimensional impact that stress can have on human behavior. In the past, several works have tried to address the problem of stress detection using different types of sensors, tasks and processing techniques [5, 6, 4].

Through our work, we target three main contributions. Firstly, we evaluate our study using the MuSE Datsaset [11], to our knowledge one of the richest databases for stress and emotion analysis in terms of stimuli and recorded modalities. One of the most valuable characteristics of MuSE compared to other available resources is that stress is not induced to the subjects through a specific task that they need to complete during data collection. In contrast, the dataset aims to capture stress as experienced by the participants in their real lives during the time of the recordings. We discuss more about the characteristics of the dataset later on in the paper on the corresponding section. Secondly, we aim to overcome the process of manually handpicking features for each individual modality by proposing and evaluating a set of different deep learning based configurations for affective modeling and stress detection. We showcase the potential of such approaches and we highlight the advantage of designing modular deep architectures that can learn unsupervised features in a task-agnostic manner, hence increasing the generalizability and applicability of pretrained components across different tasks and applications. Lastly, we propose a preliminary approach towards learning modality-agnostic representations. Different sensors introduce different limitations that can relate to variations in computational demands, sampling rate, data availability, and most importantly, a modality-based preprocessing and feature design process. Overcoming these obstacles is one of the greatest challenges in most multimodal processing tasks and our modular method demonstrates a potential solution towards that direction.

2 Related Work

Various computational methods have been proposed over the years aiming to capture and characterize human behaviors related to stress. Technologies related to sensor-based monitoring of the human body have the lion's share in this domain and different modalities and stress stimuli scenarios have been explored.

With one of the most impacting works in the area, [5] showed that multimodal monitoring can effectively capture changes in the human behavior related to stress and that specific factors such as body acceleration, intensity and duration of touch as well as the overall amount of body movement can be greatly affected when acting under stress. [1] enhanced these preliminary findings, by identifying that fluctuations in physiological features like blood volume, pulse and heart rate, may indicate significant changes in the levels of stress. The very insightful review study published by [2], emphasized on the importance of considering psychological, physiological, behavioural and contextual information when assessing stress, primarily due to the intricate implications that it can have on behavior.

Their review suggested a plethora of features, extracted from various modalities including cameras, thermal imaging, physiological indicators, environmental and life factors and several others, as important predictors of stress. In a more recent study, [6] investigated the impact that stress can have on driver behavior by monitoring some of the physiological signals indicated by the previous studies. The authors explored a series of temporal and spectral hand-crafted features and evaluated their methods using traditional machine learning approaches. All the aforementioned findings have been consistently revisited, reevaluated and most of the times reconfirmed by a series of survey studies that addressed multimodal stress detection over the last few years [21, 9, 3].

This work has been significantly inspired by the research studies of the past. However, in contrast to the works discussed above, we aim to approach multimodal stress detection using deep learning modeling. Our motivation, stems from the very inspiring results that deep learning has offered to the computer science community. In the past, very few studies explored deep learning as a tool for stress classification and feature extraction, primarily due to the limited amount of available resources. A factor that can become a very hard constrain given the excessive amounts of data that most deep learning algorithms require. Some of the most popular deep learning based studies related to stress, include the works by [14] on textual data extracted from the social media and [12] on audio data generated by actors simulating stress and non-stress behaviors.

In contrast to those techniques, we perform multi-modal processing using spatiotemporal analysis on eight different information channels with minimal data prepossessing [22] captured from 28 eight individuals. A subject set greater than all the research studies mentioned above. We explore the potentials of Recurrent Neural Networks[15, 7] and Convolutional Autoencoders [10] for learning affective representations and we do an in depth evaluation of our techniques using the MuSE dataset.

3 Dataset

MuSE is a multimodal database that has been specifically designed to address the problem of stress detection and its relation to human emotion [11]. The dataset consists of 224 recordings coming from 28 subjects who participated into two recording sessions each. All subjects were undergraduate or graduate students from different majors. The first recording session took place during a final exam period (consisdered to be a high stress period), while the second one was conducted after the exams ended (considered to be a low stress period). During each session, subjects were exposed to four different stimuli, which aimed to elicit a variation of emotional responses.

The stimuli used in each recording session were the following:

1. Neutral: Subjects were just sitting while multimodal data were being collected. No emotional stimulus was provided.

2. Question Answering (QA): Subjects were asked to answer a series of controversial questions that appeared on a screen. Questions were targeted to achieve either a positive, a negative or a neutral feeling as aftereffect.
3. Video: Subjects were asked to watch a series of emotionally provocative videos. Similarly to QA, videos were aiming to trigger a variety of emotions.
4. Monologues: After the end of each video subjects were asked to comment for 30 seconds on the video they just watched.

During all four steps in both recording sessions the following eight different streams of multimodal data were collected:

1. Thermal Imaging: Thermal imaging data of subject's face were collected during the whole period of each session.
2. RGB Closeup Video: Regular RGB video of subject's face was recorded during the whole duration of a session.
3. RGB Wideangle Video: RGB video recordigs showing a full-body view of the subject was also captured.
4. Audio: User verbal responses were recorded for the interactive sections of each recording, i.e., the QA and monologues.
5. Physiological: Four different types of physiological data were recorded using contact sensors attached to the subject's fingers and core body. The physiological signals captured were: (1) Heart rate; (2) Body temperature; (3) Skin conductance; (4) Breathing rate.
6. Text: Transcripts extracted from the QA. For the purposes of this study we did not conducted any experiments using this modality.

Table 1 summarizes the statistics of the final version of the MuSE as curated for our experiments. In this study, we consider as a sample of Stress or Non-Stress any segment that was captured during an exam or post-exam recording session respectively. Thus, data-points that belong in the same class may originate from different stimuli as long as they have been captured during the same period.

Modality	N(%)	S(%)	Total
Thermal	319 (49.9)	320 (50.1)	639
RGB Closeup	336 (51.1)	322 (48.9)	658
RGB Wideangle	336 (51.1)	322 (48.9)	658
Audio	139 (50.7)	135 (49.3)	274
Physiological	364 (49.1)	378 (50.9)	742
Total	1494 (50.3)	1477 (49.7)	2971

Table 1. Total number of samples in each class for each modality. In the case of physiological data each sample represents an instance of all four physiological signals captured, ie. Heart Rate, Body Temperature, Skin Conductance and Breathing Rate. In the parentheses we report the correspondent percentage, which is equal to the random choice accuracy. In all cases random choice is very close to 50%.

Figure 1 illustrates the distribution of samples for each subject across the two classes. Based on these data distributions we conducted all of the experiments presented later in the "Experimental Results" Section.

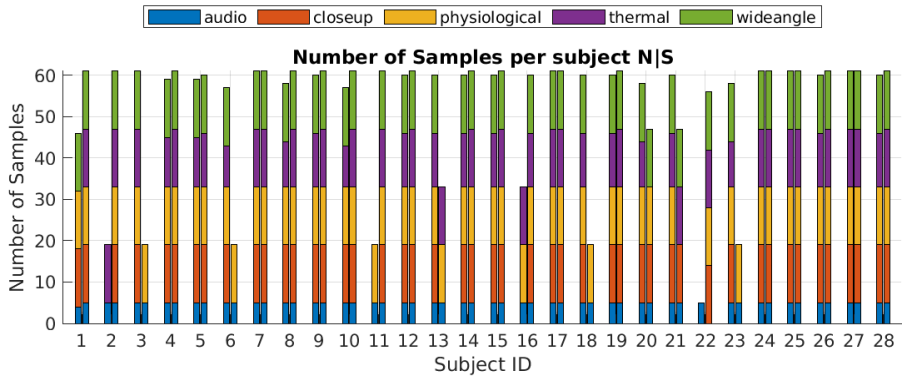

Fig. 1. Number of samples for each subject across the two classes for each modality. Left columns correspond to No-Stress (N) and right column to Stress (S).

4 Methodology

For our experiments we propose a deep-learning architecture that is based on Convolutional-Autoencoders and Recurrent Neural Networks. As briefly discussed in Section 2, Convolutional-Autoencoders (CAs) are popular for their ability to learn meaningful unsupervised feature representations by significantly shrinking the dimensionality of the original signal [23, **?**]. On the other hand, Recurrent Neural Networks have shown state-of-the-art results in a series of applications across different domains and they are mostly popular for their benefits on sequential information modeling [15]. For our implementation we used a particular recurrent unit also known as Gated Recurrent Unit (GRU) [7].

The novelty of our approach stems from the fact that we use multiple identical copies of the same architecture to model each modality individually, while applying minimal preprocessing steps on the original signals. Figure 2 illustrates the basic components of this architecture.

In addition we propose a novel approach towards modality independent multimodal representations using a modified version of our original architecture that allows weight-sharing across all the available modalities while taking into account modality dependent characteristics. For the purposes of this paper we refer to this approach as "a-modal" and we visualize it in Figure 3.

In the following subsections we will discuss in more detail the exact steps used for preprocessing each modality as well as the individual components of each architecture.

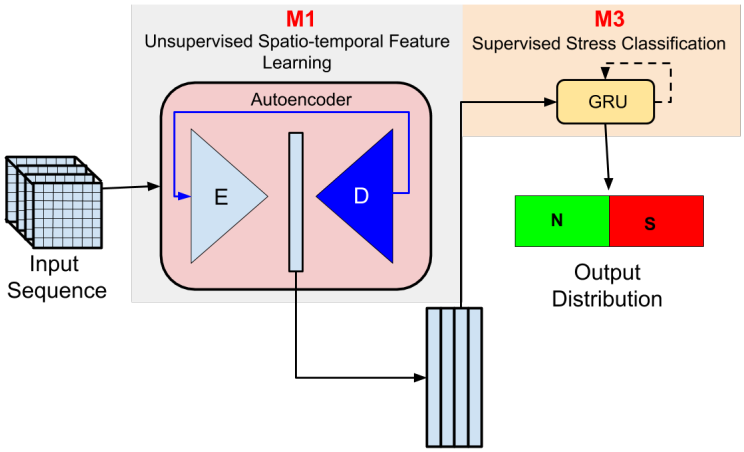

Fig. 2. Architecture for modality-dependent classification. The architecture consist of two main modules 'M1' and 'M3', which perform unsupervised feature learning and supervised stress classification respectively. The two components are trained independently. Dark-blue parts represent operations/components that are used only during training but are omitted when testing. E and D refer to Encoder and Decoder respectively, while N and S to No-Stress and Stress classes.

4.1 Modality Preprocessing

We try to minimize the computational workload of our framework by significantly simplifying the preprocessing steps applied on each modality. Below we describe the computations applied on each information signal before entering the initial encoder-unit of our deep architectures.

1. Thermal: Thermal video included in MuSE was captured in a frame-rate of 30 fps. To minimize the amount of information we clamp all temperatures between 0 and 50 degrees Celsius. Before passing the thermal video frames through the network we resize each frame to 128×128 and we convert each frame into gray scale.
2. RGB: Wide-angle & Closeup video streams were in an original frame-rate of 25 FPS. The frames from these modalities were directly re-scaled to 128×128 and converted to gray scale without any additional edits.
3. Physiological: All four physiological indicators described in Section 3 were captured in a sampling-rate of 2048 Hz. For each of the signals we extract 2 sec. windows with a 98% overlap and we compute a Fast Fourier Transform (FFT) on each of the individual segments for each of the signals. Finally, the four spectra are being stacked vertically to form a 4×4096 matrix representation for all the physiological signals combined. This representation is used as a final input to the network.
4. Audio: The audio signal is recorded at a sample rate of 44.1 kHz. Similar to the physiological signals, we extract overlapping windows of size 0.37

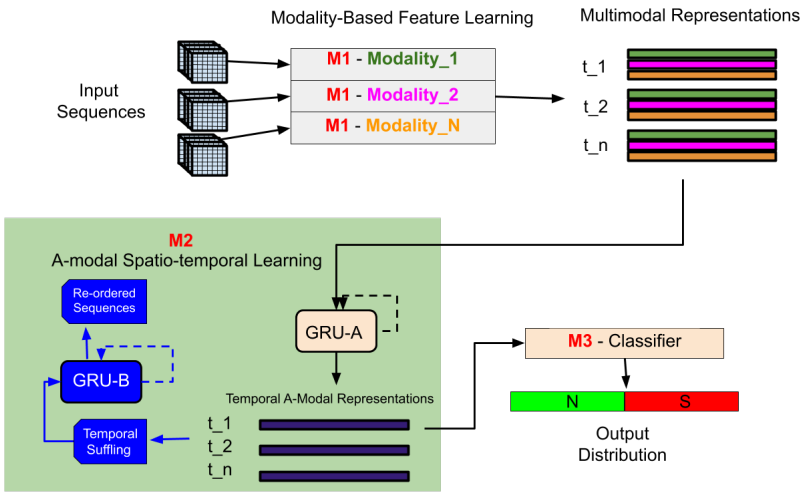

Fig. 3. The a-modal architecture. In addition to 'M1' and 'M3' this model has an extra component 'M2' that is responsible to project the multimodal representations in a new, modality-agnostic feature space, by maintaining their temporal coherence. This, happens by pretraining 'M2' on a sequence-sorting task using self-supervision. 'M1', 'M2' and 'M3' can be trained consecutively and independently. Classification happens using 'M3' in a similar manner as in Figure 2. Dark-blue parts represent operations/components that are used only during training but are omitted when testing. N and S refer to No-Stress and Stress classes.

sec. and compute the FFT on each of the windows to create a final audio representation of 1×16384, which is passed through the deep architecture. The window overlap for the audio signal is equal to 92%. This is a common way of representing audio used in previous work [18, 19]

Unsupervised Feature Learning ('M1' Module) As explained in the beginning of this section we propose two different architectures which, share some common core characteristics. The main component shared by both designs is the 'M1' module, which can be seen in detail in Figure 2. This module consists of a Convolutional-Autoencoder with 14 symmetrical convolutional layers, 7 layers for encoding and 7 for decoding. The encoding portion has kernel sizes of 3x3 with the number of filters per layer as follows: 2,4,8,16,32,64,128. The decoding section is a mirror image of the encoder. Every convolutional layer is followed by a ReLU layer except for the last encoding layer which is a sigmoid. All convolutional layers have a stride and padding of 1. We used an adaptive learning starting at 0.01. Every time the loss fails to improve for 5 epochs, the learning rate is halved. The Autoencoder is being trained independently for each modality with the objective to minimize the L1 difference between the original inputs and the output matrices generated by the decoder. After training an Autoencoder

on each modality we ignore the decoding part and use the encoder to produce modality-specific vectorized representations with a fixed size of 1×128. The Autoencoder architecture is almost identical for all modalities with the only difference being in the size of the initial input layer in order to facilitate the specific representation of each modality as discussed in the previous section. In our a-modal architecture (Figure 3), multiple copies of 'M1' are being used depending on the number of available modalities.

Learning A-modal Representations ('M2' Module) Goal of the a-modal architecture is to create a feature space that can sufficiently describe all modalities not only in the spatial but also in the temporal domain, without being restricted in the nature or the number of the available information signals.

One of the most popular obstacles in multimodal representation learning is the frame-rate miss-matching across the different signals, which makes it difficult to temporally align the various data-streams in the processing level. In our case, we try to match all modalities to the maximum frame-rate of 30 fps provided by the Thermal camera. Closeup and Wideangle RGB videos have a frame-rate of 25 fps. To "correct" the frame-rate of these two sources we simply up-sample the signal by duplicating every 5th frame of the original video. For the physiological and audio signals, we extract 30 windows per second based on the principles described previously in Section 4.1.

After fixating all modalities to the same frame-rate we use 'M1' module as shown in Figure 3 to extract a vector of 1×128 from each of them. Thus, at every frame per second we get a set of $N \times 128$ feature vectors, where N equals the number of available modalities. The main component that discriminates the original architecture of Figure 2 from the a-modal architecture is module 'M2', shown again in Figure 3. Goal of this module is to project the multimodal representations in a new, modality-agnostic feature space, by maintaining their temporal coherence.

'M2' module consists of two GRU components. GRU-A is a unidirectional RNN responsible to project the spatio-temporal, multimodal representations into the new a-modal space. To tune the parameters of GRU-A we use a another, biderectional GRU (GRU-B), that aims to solve a frame-sorting problem using the a-modal representations generated by GRU-A. This step is implemented using self-supervision and it was inspired by the work shown by [13]. Thus, no task-specific annotations are required to train 'M2'. The two components are trained together using a shared objective function that aims to optimize the sorting task by improving the quality of the projected representations. Similarly to 'M1', 'M2' was trained independently. The learning function of 'M2' is shown below:

$$L = \min_{p,P} ||p_0 - p_n|| + ||P - \hat{P}||$$

Where $\hat{P}$ is the reference temporal permutation matrix, P is the output of GRU-B and represents the predicted temporal permutation matrix, p_0 is the

output of the GRU-A over a single modality and p_n is the output of the GRU-A over all modalities.

During testing, the pretrained GRU-A is used to produce a-modal projections of the new/unknown multimodal samples, while GRU-B is omitted from the pipeline.

4.2 Stress Classification ('M3' Module)

In both the modality-dependent and modality-independent architectures, classification takes place using a "time-aware" unidirectional GRU, shown as Module-'M3' in both Figures 2 and 3. 'M3' is the only component that is trained in a fully-supervised manner.

In the first case of modality-dependent classification, 'M3' takes as input a matrix of size $h \times 128$, where h is a hyperparameter representing the temporal window on which we make classification decisions and is depended on the frame-rate of each modality. This matrix is generated by stacking consecutive vectorized representations generated by the pretrained Encoder of 'M1'. For our experiments we make classification decisions based on 20 second long overlapping windows with a 5 second step. We also perform early and late fusion experiments by combining all the available information signals. In the first case, early fusion happens by concatenating the modality-based 1D representations generated by the 'M1' Encoder. In the later case of late fusion, we vote over the available unimodal decisions using each models' individual average accuracy as a weighting factor.

In the case of a-modal classification the input to 'M3' is again a stack of feature vectors of size $h \times 128$, with the main difference being that h=600 in all scenarios, given that all modalities have a fixated frame-rate of 30 fps and that we still classify 20 seconds long windows.

5 Experimental Results

We have conducted two categories of experiments that differ on the amount of input modalities considered for the final decision making. For all our experiments we perform subject-based leave-one-out cross validation and we report the average performance metrics across all 28 subject.

Since meaningful verbal interaction was present only in parts of the recordings, specifically for the audio modality we have performed analysis only in the QA recording segments where plenty of meaningful audio samples were available. We excluded Monologues, since in most cases audio samples were very short and poor of verbal and linguistic information with very long pauses.

Table 2 illustrates the final stress classification results using only a single modality. For these experiments we deploy exclusively the architecture of Figure 2, as described in Section 4.

As it can be observed by the stability occurred across all the reported metrics, the classification results are pretty balanced between the two classes in all

modalities. A result that is in line with the balanced nature of the dataset as shown in Table 1 and that proves the ability of the general architecture of Figure 2 to capture and discriminate the valuable information in most scenarios.

However, despite the classification improvement observed compared to random choice in all cases, not all modalities were equally good on detecting stress. In particular, Closeup video and Physiological sensors showed the minimum improvement with 1.3% and 2.4% increase against random respectively, while Wideangle video was by far the best indicator of stress with an increase of 35%. The superiority of Wideangle imagining can be attributed to the fact that overall body, arm and head motion is being captured from this point of view, features known for their high correlation to stress as explained in Section 2. On the other hand, we believe that the poor performance of physiological sensors is due to the the modality prepossessing performed before passing the signals through module "M1", as our results are contradictory to most related research. In the future we would like to investigate more temporal-based or spectral-temporal combined physiological signal representations as others have done in the past, since focusing explicitly on spectral information seems to ignore very important characteristics of the individual signals. Other aspects such as signal segmentation and examination of the unsupervised features learned should also be revisited and reexamined. With respect to the Closeup video, our post-analysis revealed that the Autoencoder of module "M1" failed in the vast majority of cases to recreate facial features that could be indicative of stress. In most scenarios, the images recreated by the decoder, were lacking the presence of eyes and lips and only the head posture could be partially reproduced. We suspect that a reason for this effect might have been the variability of features present in the our training data, in combination to the limited amount of available samples. Thus, causing the Autoencoder to overfit on the background information. In the future, we would like to experiment on transfer-learning approaches by fine-tuning a pretrained Autoencoder model on facial data, since such methods have shown promising results in a variety of applications. Lastly, Thermal and Audio signals provided also noticeable improvements against random choice with 9.9% and 19.5% increase accordingly. It has to be noted that since Audio was considered only in the QA sections, the available samples were significantly limited. This emphasizes the effectiveness of the proposed method to capture impactful affective audio features without the need of vast amounts of data.

In Figure 4 we illustrate sample images as they were recreated by the the Autoencoder of "M1" for the Wideangle, Thermal and Closeup videos. It is easy to observe that the more details included in the reconstructed image the highest the performance of the individual modality. In the case of Wideangle, body postures can be depicted quite well, while in Thermal images the warmer areas of the face (a feature that we can intuitively understand that may be similar across different subjects under stress) have been satisfactorily captured. However as explained above Closeup images could not be represented efficiently.

Recordings	Modality	Acc	Pr	Rec	F1
All	Thermal	59.9	60.1	59.7	59.9
All	Wideangle	86.1	85.8	85.9	85.8
All	Closeup	52.4	53.8	52.2	53.0
All	Physiological	53.3	55.0	50.1	52.5
QA	Audio	70.2	70.4	70.4	70.3

Table 2. Results as percentages on modality-dependent experiments using the architecture of Figure 2. Audio was analyzed using only the samples available in the QA section. Acc, Pr and Rec refer to accuracy, precision and recall evaluation metrics respectively. Reported results are averaged across all users after a leave-one-out subject-based evaluation across all 28 subjects.

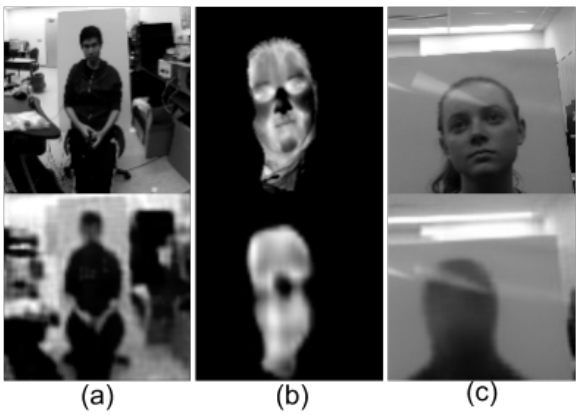

Fig. 4. Reconstructed images by the Autoencoder of module "M1". Top row corresponds to the original images and bottom one to the generated ones. Images in (a) refer to Wideangle, (b) to Thermal and (c) to Closeup videos.

Table 3 corresponds to multimodal experimental results. We report early and and late fusion results based on the architecture of Figure 2 and a-modal fusion based on the top performing modalities using the architecture of Figure 3.

Early fusion was conducted by concatenating the modality-based vectors generated by "M1" before passing them through "M3" while late fusion was a simple voting across the different decisions made by each unimodal classifier. Both early and late fusion were conducted on all the available modalities.

For a-modal fusion we perform two different experiments, one using only Audio and Wideangle on the QA recording segments (top two performing modalities) and one using Wideangle, Physiological and Thermal on all the available data (top three performing modalities). Thus, illustrating the flexibility and the potential of this approach.

As mentioned before, since valuable verbal interaction is available mostly in the QA recordings, for each fusion method we perform two experiments. One using exclusively the QA recordings ,where all five modalities where consid-

ered, and one on all the available recordings where we considered only Thermal, Wideangle, Closeup and Physiological information.

Our results indicate that late fusion of the individual decisions provided by far the best results in all cases. Early fusion could not scale up its performance when audio features where not available and showed overall inferior performance compared to the other two fusion techniques. A-modal fusion also provided relatively poor results compared to the modality-dependent late fusion approach. However, a-modal results were overall slightly better to early fusion in terms of average accuracy across the two types of experiments (QA and All recordings). Moreover a-modal fusion performed better compared to Closeup and Physiological unimodal models and provided results slightly inferior to the Thermal unimodal approach. In addition, the results provided by the a-modal approach are very stable between the two experiments (similarly to the late fusion), despite the fact that completely different modalities were used. This observation may be indicative of the stability of the learned a-modal representations, but further experimentation is needed. However, this was not the case in early fusion, since the two experiments (QA vs All) had a 5.6% difference in performance despite the fact that the majority of the modalities were the same. These results prove the ability of the a-modal method to learn robust, modality-agnostic representations that carry and combine affective knowledge from all the available resources. Our findings indicate that there is obviously a long way to go until models of general affective awareness become a reality. However, they highlight the possibilities of such methods and motivate us towards investigating this topic further.

Recordings	Fusion	Acc	Pr	Rec	F1	Avg Acc
All	Early	53.0	52.1	51.7	51.9	54.6
QA	Early	54.7	61.8	53.4	57.3	
All	Late	87.8	87.2	87.9	87.6	88.5
QA	Late	89.4	89.3	89.4	89.3	
All	A-modal	56.1	55.7	57.3	56.6	56.8
QA	A-modal	57.3	58.2	55.7	56.9	

Table 3. Results as percentages on multimodal experiments. We perform three types of such experiments; early and late fusion using the modality-dependent architecture of Figure 2 and a-modal fusion based on the architecture of Figure 3. Each method was tested both on all recording parts of MuSE, using all modalities except audio and on the QA alone by including audio. For the a-modal experiments we evaluate only using the top performing modalities as shown in Table 2. In particular, when testing on all the recording parts, Physiological, Wideangle and Thermal signals were used while on QA we evaluated using Wideangle and Audio. Acc, Pr and Rec refer to accuracy, precision and recall. Results are averaged across all users after a leave-one-out subject-based evaluation across all 28 subjects. Avg Acc refers to the average accuracy between the two experiments conducted on the different recording segments.

6 Conclusions & Future Work

In this paper we investigated the abilities of deep-learning methods on producing affective multimodal representations related to stress. We proposed a modular approach that enables learning unsupervised spatial or spatio-temporal features, depending on the way that the different modules are combined. We showed how each module can be trained and reused independently and how different combinations of the modules can lead to different ways of combining modalities, each coming with its own advantages and disadvantages.

In particular we demonstrated an architecture (Figure 2) able to learn spatial modality-dependent representations in a modality agnostic way and we evaluated the abilities of each information channel to capture signs of stress. Additionally, we proposed a variation of this original architecture (Figure 3) able to produce modality-independent representations, by operating on an arbitrary number of input signals that can be highly unrelated with each other but very informative towards understanding the targeted task; in this case the detection of stress.

One of the main assets of the proposed method is its ability to provide promising results across all the evaluated experiments by minimizing the preprocessing steps of all the available signals and by completely avoiding manual feature engineering. The presented results showcase that deep-learning methods can produce rich affective representations related to stress, despite the relatively limited amount of data. Moreover, they show that they can function as mechanisms to process, extract and combine information coming from multiple resources without the need of explicitly tailoring each classifier on the characteristics of each individual modality. These findings motivate us towards researching these topics in greater depth.

In the future we would like to investigate alternative approaches of representing the different modalities before processing them through the deep architectures as we believe that it can highly impact the performance of the model. However, our priority is to do so without compromising the minimal computational preprocessing cost as discussed in this paper. Furthermore, we plan to apply our methods on other applications in the spectrum of affective computing such as alertness, fatigue and deception detection. Finally, we would like to investigate alternative architectures, that can lead to improved results both in terms of classification and computational performance.

Acknowledgments

This material is based in part upon work supported by the Toyota Research Institute (TRI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of TRI or any other Toyota entity.

References

1. Aigrain, J., Spodenkiewicz, M., Dubuiss, S., Detyniecki, M., Cohen, D., Chetouani, M.: Multimodal stress detection from multiple assessments. IEEE Transactions on Affective Computing **9**(4), 491–506 (2016)
2. Alberdi, A., Aztiria, A., Basarab, A.: Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. Journal of biomedical informatics **59**, 49–75 (2016)
3. Can, Y.S., Arnrich, B., Ersoy, C.: Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. Journal of biomedical informatics p. 103139 (2019)
4. Can, Y.S., Chalabianloo, N., Ekiz, D., Ersoy, C.: Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. Sensors **19**(8), 1849 (2019)
5. Carneiro, D., Castillo, J.C., Novais, P., Fernández-Caballero, A., Neves, J.: Multimodal behavioral analysis for non-invasive stress detection. Expert Systems with Applications **39**(18), 13376–13389 (2012)
6. Chen, L.l., Zhao, Y., Ye, P.f., Zhang, J., Zou, J.z.: Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. Expert Systems with Applications **85**, 279–291 (2017)
7. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
8. Dobson, H., Smith, R.: What is stress, and how does it affect reproduction? Animal reproduction science **60**, 743–752 (2000)
9. Greene, S., Thapliyal, H., Caban-Holt, A.: A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health. IEEE Consumer Electronics Magazine **5**(4), 44–56 (2016)
10. Guo, X., Liu, X., Zhu, E., Yin, J.: Deep clustering with convolutional autoencoders. In: International Conference on Neural Information Processing. pp. 373–382. Springer (2017)
11. Jaiswal, M., Aldeneh, Z., Bara, C.P., Luo, Y., Burzo, M., Mihalcea, R., Provost, E.M.: Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7415–7419. IEEE (2019)
12. Lane, N.D., Georgiev, P., Qendro, L.: Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 283–294. ACM (2015)
13. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 667–676 (2017)
14. Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Cai, L., Feng, L.: User-level psychological stress detection from social media using deep neural network. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 507–516. ACM (2014)
15. Mandic, D.P., Chambers, J.: Recurrent neural networks for prediction: learning algorithms, architectures and stability. John Wiley & Sons, Inc. (2001)
16. McEwen, B.S.: Protective and damaging effects of stress mediators. New England journal of medicine **338**(3), 171–179 (1998)

17. Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective multimodal human-computer interaction. In: Proceedings of the 13th annual ACM international conference on Multimedia. pp. 669–676. ACM (2005)
18. Papakostas, M., Giannakopoulos, T.: Speech-music discrimination using deep visual feature extractors. Expert Systems with Applications **114**, 334–344 (2018)
19. Papakostas, M., Spyrou, E., Giannakopoulos, T., Siantikos, G., Sgouropoulos, D., Mylonas, P., Makedon, F.: Deep visual attributes vs. hand-crafted audio features on multidomain speech emotion recognition. Computation **5**(2), 26 (2017)
20. Picard, R.W.: Affective computing. MIT press (2000)
21. Sharma, N., Gedeon, T.: Objective measures, sensors and computational techniques for stress recognition and classification: A survey. Computer methods and programs in biomedicine **108**(3), 1287–1301 (2012)
22. Tan, S., Guo, A., Ma, J., Ren, S.: Personal affective trait computing using multiple data sources. In: 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). pp. 66–73. IEEE (2019)
23. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1274–1283 (2017)