Table 2. Detailed performance metrics of proposed system

Dataset	Without Edge Learning			With Edge Learning		
	Precision	Recall	F-1 Score	Precision	Recall	F-1 Score
DMD	0.89	0.88	0.88	0.92	0.92	0.92
YawDD	0.85	0.82	0.83	0.87	0.85	0.86

After this, in order to check the on-board learning capability of SNN on Akida NSoC, we have performed *Edge Learning* for which the last classification layer was replaced with an *edge learnable layer*. Akida provides flexibility to train only this newly added layer and also allows to add more than one neurons per class for on-board learning. In this case, we have used 100 neurons for each of the classes for on-board training of this last layer. To note, the model is not going to learn new classes at present, but we intend to check whether the overall performance of the model improves after this on-device training on new unseen subject data. This modified edge-learnable model was trained using *Akida unsupervised learning function* for optimal performance on Akida hardware with a small subset of (approx. 20%) unseen test data for a single epoch. The results obtained are reported below.

Stage I - Results: Table 1 summarizes the result so obtained in Stage I experiments. As mentioned above, the model has been tested with and without edge learning capability. Although the accuracy of the model without edge learning is 88.2% and 88.85% for DMD and YawDD datasets respectively, it increases to 92.01% and 90.05% after applying edge learning. Thus adding the *Edge Learning Layer* and re-training with 20% of test data helps improve the model performance. Moreover, this performance is at par with state of the art deep learning models. Our model can infer around 21 images per second i.e. taking 46.52 ms per image while consuming at max. 16.21 mJ per frame.

Table 2 reports the changes in classification metrics for both the datasets before and after applying edge learning. For DMD dataset, the precision increases from 0.89 to 0.92, recall increases from 0.88 to 0.92 and F-1 score increases from 0.88 to 0.92 before and after applying edge learning respectively. In case of YawDD dataset, the precision increases from 0.85 to 0.87, recall increases from 0.82 to 0.85 and F-1 score increases from 0.83 to 0.86 before and after applying edge learning respectively.

Stage II - Training: Encouraging results related to *Edge Learning* of Stage-I experiments inspired us to study further the capability of the model in learning new classes via on-board learning. DMD dataset, having four classes, has been used for Stage II experiments. In first set of experiments under Stage II, model was initially trained on three classes from DMD dataset and then deployed on AKD1000 NSoC. The last layer was replaced by an *Edge Learnable Layer*, like before, consisting of 100 neurons per class with a target to learn all four classes of DMD. The model was then trained (only the last layer) for one epoch using

Akida unsupervised learning on 150 samples of each class (which is equivalent to roughly 5 s of video). We have taken all possible combinations of “three class training & one class learning at edge”. In the second set of Stage II experiment, similar activities were repeated with initial training on two classes and targeting to learn rest two classes of DMD at edge. Here again, all possible combinations of “two class training &two class learning at edge” were considered. 150 samples per class was used in this case too. The results of these Stage II experiments are reported below.

Table 3. Performance with 3 class model extended to 4 class on Akida

3 Classes for which the model is initially trained	SNN Accuracy for 3 classes	1 new class added during Edge Learning	SNN accuracy for 4 classes	Accuracy drop w.r.t Stage I model
Eyes Closed, Eyes Open, Yawning with hand	87.26%	Yawning without hand	79.82%	8.38%
Eyes Closed, Eyes Open, Yawning without hand	89.14%	Yawning with hand	82.24%	5.96%
Eyes Closed, Yawning with hand, Yawning without hand	88.58%	Eyes Open	88.43%	-0.23%
Eyes Open, Yawning with hand, Yawning without hand	91.56%	Eyes Closed	81.76%	6.44%

Stage II - Results: Table 3 depicts the performance of model being trained on three class and one additional class being learnt during on-board *Edge Learning*. Clearly, there is a drop in the accuracy before and after on-board training because of the introduction of new class at edge. As the training of new class is taking place over the quantized weights of the model, the weight modifications are constrained to the quantization bit limit, that too for the last layer only. Moreover, we are training only for one epoch. These reasons are not allowing the model to learn the new class to fullest, thereby resulting into a drop in accuracy. However, the drop remains within an allowable range (on average 7%), and in some cases, the accuracy drop is negligible (row 3 of Table 3). This opens up the avenue to further enquire the effect of increased number of epochs on edge-learning.

In Table 4, the performance of model being trained on two class and learning two additional class at edge is shown. As expected, the accuracy drop in this case is more compared to Table 3 (min drop: 21%, max drop: 40%, average drop: 29%) for the reasons mentioned above. The change in accuracy drop values from Table 3 to Table 4 signifies that more training samples and epochs may be required by the model to improve upon its performance.

Table 4. Performance with 2 class model extended to 4 class on Akida

2 Classes for which the model is initially trained	SNN Accuracy for 2 classes	2 new classes added during Edge Learning	SNN accuracy for 4 classes	Accuracy drop w.r.t Stage I model
Eyes closed, Eyes open	90.73%	Yawning with hand, Yawning without hand	56.63%	31.57%
Eyes closed, Yawning with hand	88.76%	Eyes open, Yawning without hand	60.64%	27.56%
Eyes closed, Yawning without hand	94.94%	Eyes open, Yawning with hand	66.32%	21.88%
Eyes open, Yawning with hand	92.41%	Eyes closed, Yawning without hand	71.36%	16.84%
Eyes open, Yawning without hand	94.89%	Eyes closed, Yawning with hand	70.56%	17.64%
Yawning with hand, Yawning without hand	95.66%	Eyes closed, Eyes open	54.08%	34.12%

The promising results of on-board *Edge Learning* opens up avenue for personalization. The ability to learn new classes (i.e. new activities carrying personal signature) on-board with very less data will allow a driver/car-owner to personalize the model on the basis of his/her personal habits and features and further improve the monitoring system.

The power and latency figures, as reported in Table 1 are promising and suggest that our system can be used for real time needs in battery driven cars where power is scarce.

6 Conclusion and Future Works

Driver drowsiness is an important contributing factor to the number of car accidents and insurance claims around the world. In this paper, we have proposed one in-car AI enabled driver drowsiness detection system using latest AI paradigm called Neuromorphic computing and SNN. Our solution achieves quite high accuracy and promises real-time response with all the computation being done using very low power on Akida Neuromorphic SoC. We have further established that effective on-device learning on unseen data or new class of activities is possible using our system thereby opening up the opportunity of personalization of the system. In future, we aim to focus on performing the same task using real-time streaming data from a DVS camera, instead of the RGB camera. However, the integration of the proposed system with car circuitry is a challenge that needs to be addressed.

References

1. Fatigued driver national safety council. https://www.nsc.org/road/safety-topics/fatigued-driver
2. Driver-monitoring. https://disa.com/dot-transportation-compliance/driver-monitoring
3. Hussein, M.K., Salman, T.M., Miry, A.H., Subhi, M.A.: Driver drowsiness detection techniques: a survey. In: 2021 1st Babylon International Conference on Information Technology and Science (BICITS), pp. 45–51 (2021)
4. Lawoyin, S., Fei, D.-Y., Bai, O.: Accelerometer-based steering-wheel movement monitoring for drowsy-driving detection. In: Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, vol. 229, pp. 163–173 (2014)
5. Chowdhury, A., Shankaran, R., Kavakli, M., Haque, Md.M.: Sensor applications and physiological features in drivers' drowsiness detection: a review. IEEE Sens. J. **18**(8), 3055–3067 (2018)
6. Iwamoto, H., Hori, K., Fujiwara, K., Kano, M.: Real-driving-implementable drowsy driving detection method using heart rate variability based on long short-term memory and autoencoder. In: IFAC-PapersOnLine, vol. **54**, no. 15, pp. 526–531 (2021). 11th IFAC Symposium on Biological and Medical Systems BMS 2021
7. Davies, M., et al.: Loihi: a neuromorphic manycore processor with on-chip learning. IEEE Micro **38**(1), 82–99 (2018)
8. Brainchip akida neuromorphic soc. https://doc.brainchipinc.com/
9. Ortega, J.D., et al.: DMD: a large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In: Bartoli, A., Fusiello, A. (eds.) ECCV 2020. LNCS, vol. 12538, pp. 387–405. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-66823-5_23
10. Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., Hariri, B.: Yawdd: a yawning detection dataset, pp. 24–28 (2014)
11. Lawoyin, S.A., Fei, D.-Y., Bai, O., et al.: A novel application of inertial measurement units (IMUS) as vehicular technologies for drowsy driving detection via steering wheel movement. Open J. Saf. Sci. Technol. **4**(04), 166 (2014)
12. Fujiwara, K., et al.: Heart rate variability-based driver drowsiness detection and its validation with EEG. IEEE Trans. Biomed. Eng. **66**(6), 1769–1778 (2018)
13. Ahn, S., Nguyen, T., Jang, H., Kim, J.G., Jun, S.C.: Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous EEG, ECG, and FNIRS data. Front. Hum. Neurosci. **10**, 219 (2016)
14. Awais, M., Badruddin, N., Drieberg, M.: A hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability. Sensors **17**(9), 1991 (2017)
15. Sekar, K., Thileeban, R., et al.: Drowsiness and real-time road condition detection using heart rate sensor, accelerometer and gyroscope. Int. J. Comput. Digit. Syst. (2022)
16. Weng, C.-H., Lai, Y.-H., Lai, S.-H.: Driver drowsiness detection via a hierarchical temporal deep belief network. In: Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13, pp. 117–133. Springer (2017)

17. Park, S., Pan, F., Kang, S., Yoo, C.D.: Driver drowsiness detection system based on feature representation learning using various deep networks. In: Asian Conference on Computer Vision, pp. 154–164. Springer (2016)
18. Mandal, B., Li, L., Wang, G.S., Lin, J.: Towards detection of bus driver fatigue based on robust visual analysis of eye state. IEEE Trans. Intell. Transp. Syst. **18**(3), 545–557 (2016)
19. Lyu, J., Yuan, Z., Chen, D.: Long-term multi-granularity deep framework for driver drowsiness detection. arXiv preprint arXiv:1801.02325 (2018)
20. Kang, N., et al.: Driver drowsiness detection based on 3D convolution neural network with optimized window size. In: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), pp. 425–428 (2022)
21. Salman, R.M., Rashid, M., Roy, R., Ahsan, Md.M., Siddique, Z.:Driver drowsiness detection using ensemble convolutional neural networks on YAWDD. arXiv preprint arXiv:2112.10298 (2021)
22. Chen, K., Zhu, T., Li, S., Shi, Y.: Facial keypoint-based segment-level driver yawning detection by graph-temporal convolutional neural network modeling. Authorea Preprints (2023)
23. Cañas, P., Ortega, J.D., Nieto, M., Otaegui, O.: Detection of distraction-related actions on DMD: an image and a video-based approach comparison. In: VISIGRAPP (5: VISAPP), pp. 458–465 (2021)
24. Lakhani, S.: Applying spatiotemporal attention to identify distracted and drowsy driving with vision transformers. arXiv preprint arXiv:2207.12148 (2022)
25. Lamaazi, H., Alqassab, A., Fadul, R.A., Mizouni, R.: Smart edge-based driver drowsiness detection in mobile crowdsourcing. IEEE Access **11**, 21863–21872 (2023)
26. Li, X., Xia, J., Cao, L., Zhang, G., Feng, X.: Driver fatigue detection based on convolutional neural network and face alignment for edge computing device. In: Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, vol. 235, no. 10–11, pp. 2699–2711 (2021)
27. Rahman, A., Hriday, M., Khan, R.: Computer vision-based approach to detect fatigue driving and face mask for edge computing device. Heliyon **8**(10), e11204 (2022)
28. Becattini, F., Berlincioni, L., Cultrera, L., Bimbo, A.D.: Neuromorphic face analysis: a survey. arXiv preprint arXiv:2402.11631 (2024)
29. Berlincioni, L., et al.: Neuromorphic event-based facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4109–4119 (2023)
30. Becattini, F., Cultrera, L., Berlincioni, L., Ferrari, C., Leonardo, A., Del Bimbo, A.: Neuromorphic facial analysis with cross-modal supervision. arXiv preprint arXiv:2409.10213 (2024)
31. Kielty, P., Dilmaghani, M.S., Ryan, C., Lemley, R., Corcoran, P.: Neuromorphic sensing for yawn detection in driver drowsiness. In: Fifteenth International Conference on Machine Vision (ICMV 2022), vol. 12701, pp. 287–294. SPIE (2023)
32. Chen, G., Hong, L., Dong, J., Liu, P., Conradt, J., Knoll, A.: EDDD: event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor. IEEE Sens. J. **20**(11), 6170–6181 (2020)
33. Tavanaei, A., Ghodrati, M., Kheradpisheh, S.R., Masquelier, T., Maida, A.: Deep learning in spiking neural networks. Neural Netw. **111**, 47–63 (2019)
34. Gigie, A., George, A.M., Kumar, A.A., Dey, S., Pal, A.: Stereogest-SNN: robust gesture detection with stereo acoustic setup using spiking neural networks. In: 2023 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–4 (2023)

35. Viale, A., Marchisio, A., Martina, M., Masera, G., Shafique, M.: CarSNN: an efficient spiking neural network for event-based autonomous cars on the loihi neuromorphic research processor. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–10. IEEE (2021)
36. Kadway, C., Dey, S., Mukherjee, A., Pal, A., Bézard, G.: Low power & low latency cloud cover detection in small satellites using on-board neuromorphic processors. In: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2023)
37. Kahali, S., Dey, S., Kadway, C., Mukherjee, A., Pal, A., Suri, M.: Low-power lossless image compression on small satellite edge using spiking neural network. In: 2023 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2023)
38. Hunsberger, E., Eliasmith, C.: Spiking deep networks with LIF neurons. arXiv preprint arXiv:1510.08829 (2015)
39. Dan, Y., Poo, M.: Spike timing-dependent plasticity of neural circuits. Neuron **44**(1), 23–30 (2004)
40. Akopyan, F., et al.: Truenorth: design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **34**(10), 1537–1557 (2015)
41. Mayr, C., Hoeppner, S., Furber, S.: Spinnaker 2: a 10 million core processor system for brain simulation and machine learning. arXiv preprint arXiv:1911.02385 (2019)
42. Frenkel, C., Lefebvre, M., Legat, J.-D., Bol, D.: A 0.086-mm 212.7-pj/sop 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28-nm CMOS. IEEE Trans. Biomed. Circuits Syst. **13**(1), 145–158 (2018)
43. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

Beyond RGB: Tri-Modal Microexpression Recognition with RGB, Thermal, and Event Data

Mira Adra[1,2], Nelida Mirabet-Herranz[2](✉), and Jean-Luc Dugelay[2]

[1] GTD International, 2 Rue Giotto, 31520 Ramonville-Saint-Agne, France
mira.adra@gtd.eu
[2] EURECOM, 450 Route des Chappes, 06410 Biot, France
{mirabet,dugelay}@eurecom.fr

Abstract. Facial Emotion recognition (FER) is an extensively studied computer vision task that aims at identifying and categorizing emotional expressions depicted on a human face, such as anger, fear, or happiness. Due to the subjective nature of feelings, deep learning models may struggle to learn implicit information about a person's emotions, leading to inaccuracies in existing methods. In this work, we aim to estimate microexpressions-small facial movements that can indicate underlying feelings, as described in the Facial Action Coding System (FACS)-from face videos, as these facial movements provide explicit information that is more easily perceivable by deep learning architectures. Furthermore, despite the evolution of FER technologies driven by advancements in neural network architectures and the exploration of new sensing technologies, there is a significant shortage of datasets that leverage these emerging modalities, which limits the progress of research in this field. In our study, we aim to explore and compare the feasibility of using different input data modalities, visible, thermal, and event, as training and testing data for a CNN baseline network by presenting a pioneering dataset that integrates these three modalities, each annotated with detailed Facial Action Units (FAUs) present in the FACS. Our proposed Visible, Event, and Thermal Face Dataset for Micro Expression Recognition (VETEX) containing 2506 face videos is available upon request.

Keywords: Event Data · Thermal Spectra · Face Dataset · Microexpression · Facial Emotion Recognition · Tri-modal dataset

1 Introduction

Face videos are nowadays a key element in many applications, ranging from automatic face recognition-currently one of the most active research areas in computer vision -to soft biometric prediction and health information estimation [20]. In addition, human faces reveal information about a person's emotional status, which has driven researchers to explore the possibility of automatically

S. Palaiahnakote et al. (Eds.): ICPR 2024, LNCS 15615, pp. 311–324, 2025.
https://doi.org/10.1007/978-3-031-87660-8_23

detecting those emotions. Facial Emotion Recognition (FER) technologies aim to detect human feelings from face videos, typically using computer vision and deep learning architectures. However, several studies have highlighted that measuring emotions can be challenging due to the metaphysical and personal nature of feelings [14]. Indeed, the authors of state-of-the-art datasets for FER have pointed out the difficulty of annotating data, as subjects often report different emotions than those the authors intended to convey. This reinforces the need for double annotation-one considering the user's labeling and another following their a priori video-emotion assignment [6]. A more objective component can be found in microexpressions, which are subtle and fast movements, sometimes performed involuntarily. The fastest of them have been reported to manifest between 1/25 and 1/5 of a second [5]. Furthermore, certain microexpressions such as smiling or frowning, are also defined as one or a combination of several Facial Action Units (FAUs) and have been linked to emotions in the official Facial Action Coding System (FACS). Therefore, in this work, we propose that the FER problem can be approached more objectively by detecting FAUs, thereby eliminating the subjective component of feelings.

FER models have traditionally based their estimations on RGB videos. Despite these networks reaching a significant level of maturity with practical success [14], deep learning approaches based on visible spectrum images are affected by compromising factors such as occlusion and illumination changes [20]. In addition, traditional cameras have a low frame rate and dynamic range, which may be a barrier to human expression understanding [5]. RGB cameras, which typically operate at a maximum of 25/30 frames per second (fps), inherently struggle to capture microexpressions that manifest in short timespans of up to 1/25 of a second and might face great difficulties with FAUs recognition.

Various types of sensors have been explored in FER, including depth and 3D cameras [9], event-based data [6] and thermal imaging [15]. Event cameras, which are bio-inspired sensors, differ from traditional cameras by producing asynchronous events at individual pixels where illumination changes occur, rather than generating streams of synchronous frames [6] significantly reducing motion blur and showing higher dynamic range. They offer several advantages: extremely high temporal resolution and low latency (both in the microsecond range), a very high dynamic range (140 dB compared to 60 dB in standard cameras), and low power consumption [11]. Besides, event data representations have been highlighted in the literature as intrinsically protected data with a heightened level of security which is a critical advantage for high-security applications [3]. Furthermore, research has demonstrated how thermal imaging can be superior to visible imaging under challenging conditions such as the presence of smoke, dust, and the absence of light sources [10]. Thermal imagery works by detecting electromagnetic radiation in the medium-wave infrared (MWIR, $3 - 8\mu\text{m}$) and long-wave infrared (LWIR, $8 - 15\mu\text{m}$) spectra [23], where skin heat is detected. This capability allows thermal images to effectively handle low illumination and certain types of occlusions. Besides, event data representations have been highlighted in the literature as intrinsically protected data with a heightened level of security and efficiency in processing, which is a critical advantage for battery-powered devices or high-security applications [3].

However, despite the promising future of event and thermal input data in many applications, including preliminary studies that have shown their suitability for FER, emotion recognition through thermal and event-based videos remains a problem not widely addressed in the literature due to a lack of data. In the case of event-based data, several attempts have been made to generate synthetic event-based datasets to address this data shortage [12]. Nevertheless, no work has directly compared these three modalities, and no dataset allows for a fair comparison under similar conditions. In addition, AI-based models heavily depend on larger volumes of data for their training, and the list of face datasets in spectra other than RGB is limited. Therefore, in this article, we present the Visible, Events, and Thermal Face Dataset for Micro Expression Recognition (VETEX), the first release of a RGB, thermal, and event tri-modal dataset. To advance towards more accurate FER models and because we believe in the potential of alternative imagery compared to RGB, our main contributions are as follows:

- We present our VETEX Face Dataset, which includes 2,506 videos from 20 different subjects, totaling approximately 2.75 h of video per modality, suitable for various facial processing tasks, including FER;
- We propose the first study, to the authors' knowledge, that compares the potential of RGB, event, and thermal data for microexpression estimation using a baseline 3D CNN architecture;
- We evaluate the suitability of different input data under various illumination conditions: studio lights and no artificial light sources, resulting in natural light videos that might be poorly illuminated.

The rest of the paper is organized as follows: Sect. 2 presents advancements in the field of FER, and lists existing datasets containing thermal and event-based face videos. In Sect. 3, we provide a detailed presentation of our newly collected VETEX Face Dataset. Section 4 presents a comprehensive description of the methodology used in our experiments, as well as the experimental results of our data comparison for microexpression estimation, including a study on the impact of different illumination conditions. Finally, Sect. 5 summarizes the article and concludes with future directions for our work.

The VETEX Face Dataset is publicly available upon request.

2 Related Work

Deep learning-based FER systems are traditionally trained on datasets acquired in the visible domain or, more recently, with data in the thermal spectrum or event-based data. In this section, we present existing face datasets containing thermal and event data, besides various studies focused on FER that have considered these input data modalities.

2.1 Face Emotion Recognition

FER is a technology that analyses facial expressions from static images and videos to estimate information about a person's emotional state. Traditionally, seven emotions are targeted: happiness, sadness, anger, surprise, fear, disgust, and neutrality [28]. FER has played a significant role in cognitive psychology research, and numerous studies have focused on automated FER due to its practical significance in crowd emotion monitoring [27], driver safety assistance [7], and human-computer interactions [8].

Recent advancements in Facial Emotion Recognition (FER) have extended beyond the visible spectrum to explore the potential of thermal and event-based data, offering new solutions for detecting microexpressions under challenging conditions. One interesting research in the thermal domain combines gait information from the visible spectrum with facial data from thermal imaging, improving emotion recognition through the integration of body movement cues [16]. In another paper, Wang et al. [25] proposed a visible-thermal facial expression database and conducted experiments to analyze the relationship between facial temperature and emotion. More recently, Nguyen et al. [21] introduced a new dataset that enhances the understanding of emotional intensity by categorizing each emotion into three levels: low, medium, and high.

Event-based cameras, known for their high temporal resolution and low latency, have also gained attention in FER. Barchid et al. [4] established the first application of event cameras for FER using synthetic event data, leveraging Spiking Neural Networks to surpass traditional visible domain methods. Furthermore, Berlincioni et al. introduce the NEFER dataset [6], having visible and event data pairs, that showcases the effectiveness of event data in capturing rapid facial microexpressions that are often missed by conventional cameras.

2.2 Existing Relevant Datasets

Microexpression recognition has become an increasingly important study area within emotion recognition research. However, the available datasets remain limited, especially when considering multimodal approaches. While several datasets have been developed for FER across different modalities, only a few focus specifically on microexpressions. Moreover, many facial expression datasets have centered on the visible spectrum. RGB datasets, such as CAS(ME)^2 [22], CK+ [18], and JAFFE [19], are the most commonly used for microexpression analysis. Despite the abundance of RGB datasets, they are inherently limited by sensitivity to lighting conditions and occlusions. To address these challenges, recent research has started exploring alternative modalities, such as thermal and event-based data, which offer more robust emotion detection capabilities across diverse environments.

The LVT Face Dataset [20] expanded the scope by introducing both RGB and thermal data, enabling the study of facial biometrics across different modalities and exploring how fusing these modalities can enhance performance. In the field of FER, two of the earliest and most commonly used datasets were the IRIS [1] and NIST-Equinox [2] datasets, which offered RGB-thermal data collected under varied lighting conditions and head positions. As the FER got more popular, richer datasets were proposed like the NVIE dataset [25] containing 215 subjects, each displaying six expressions, and most recently, the KTFEv2 dataset [21] which comprises seven emotions induced by watching video clips on a screen. More recent contributions include the release of the NEFER dataset [6], which introduced both RGB and event data, enabling comparative studies between these two modalities. Additionally, NEFER is well-suited for face detection tasks due to its inclusion of bounding boxes and landmark annotations. However, like most other datasets, NEFER does not include thermal data, leaving a gap in fully exploring the advantages of a tri-modal approach.

Table 1 compares relevant face datasets based on key attributes such as the number of videos, users, modalities (RGB, Thermal, Event), lighting conditions, landmarks, and annotations. While numerous RGB-only datasets have been extensively studied, they are not included here due to their abundance [17], allowing us to concentrate on datasets that explore alternative modalities. Consequently, the table highlights the unique contribution of our proposed VETEX dataset. To our knowledge, it is the first dataset to offer a tri-modal approach (RGB, Thermal, and Event) in the field of FER. Our dataset, VETEX, marks a significant advancement as the first tri-modal microexpression dataset, incorporating RGB, thermal, and event-based data. Unlike previous datasets, VETEX is also annotated with microexpressions composed by one or more Facial Action Units (FAUs) rather than direct emotional labels. This allows for a more granular analysis of facial muscle movements, which can be mapped to emotions, providing a richer resource for microexpression recognition research. Additionally, our dataset includes data collected under various lighting conditions and from participants both with and without glasses. This diversity enables comprehensive testing of the dataset's robustness against challenging scenarios like low-light environments and occlusions, further enhancing its utility in real-world applications.

3 Dataset Description

In this section, we first introduce the recording setup of the dataset and the characteristics of the acquisition devices. We then detail the data collection protocol and present the final composition of the dataset. Additionally, we provide visual examples of frames from the different modalities included in the VETEX dataset.

3.1 Acquisition Material

To create a comprehensive multi-modal microexpression dataset suited for comparing different data input modalities, we simultaneously collected three types of facial data: RGB, thermal, and event data. Additionally, unlike other existing FER-oriented datasets, we aimed to verify generalization under different lighting

Table 1. Comparison of relevant face datasets considering RGB, event and/or thermal modalities. The table provides an overview of datasets in terms of year, modality, number of videos, users, and annotations. *These datasets are composed of frames, not videos.

Year	Dataset	# Videos	# Users	Modality			Light Conditions	Landmarks	Annotations
				RGB	TH	EV			
-	IRIS [1]	4228*	30	✓	✓	×	✓	×	Emotions
2007	NIST [2]	1919*	600	✓	✓	×	✓	×	Emotions
2010	NVIE [25]	-	215	✓	✓	×	✓	×	Emotions
2022	DFME [29]	10,045	97	✓	×	×	×	✓	Emotions
2018	TFAD [15]	2500*	90	×	✓	×	✓	✓	Landmarks/MicroExp
2022	DFME [29]	10,045	97	✓	×	×	×	✓	Emotions
2022	Becattini et al. [5]	455	25	✓	×	✓	×	×	Pos/Neg/Neutral
2023	LVT [20]	416	52	✓	✓	×	✓	×	Biometrics/eHealth
2023	NEFER [6]	609	29	✓	×	✓	×	✓	Emotions
2023	KTFEv2 [21]	1120	30	✓	✓	×	×	×	Emotions
2024	VETEX (Ours)	2506	30	✓	✓	✓	✓	×	MicroExp

conditions. Therefore, we captured videos in two different scenarios: with studio lights ensuring good illumination and under natural light conditions where the face might not be well illuminated.

The visible and thermal facial data were obtained using the dual sensor of the FLIR Duo R camera, developed by FLIR Systems. This camera is specifically designed to capture visible and thermal images simultaneously, providing precise spatial and temporal alignment for accurate data pairing. The visible and thermal sensors of this camera consist of a CCD sensor with a pixel resolution of 1920×1080 and an uncooled VOx microbolometer with a pixel resolution of 640×512, respectively.

For event data, we employ the DAVIS346 event camera with a frame size of 346×260 due to its high temporal resolution and low latency, which are critical for detecting rapid microexpressions. This camera also features a high dynamic range of 120 dB, allowing it to perform well under various lighting conditions and capture subtle aspects of microexpressions that might go undetected in other modalities (Fig. 1).

The image and video acquisition took place in an indoor environment with the ambient temperature set to 25 °C. To control the lighting conditions during data acquisition, we used two studio lights placed symmetrically on either side of the setup, securing consistent illumination on the face and enhancing the visibility of facial features. The setup included a white wall as a background, a chair positioned at a fixed distance of 0.25 m from the cameras, and a high desk to guarantee that both cameras were securely positioned, fixed, and aligned during recording. This arrangement minimized movement artifacts and ensured that the captured data was of high quality and that the faces were centered in

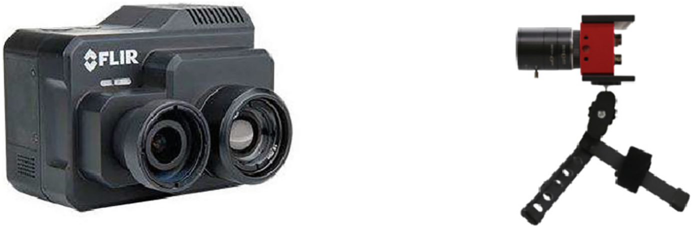

Fig. 1. Flir Duo R camera (left) and DAVIS346 event camera (right).

the frame of both cameras, facilitating accurate microexpression analysis across the three modalities.

3.2 Collection Protocol

Each of the 20 volunteers participated in one acquisition session. Before the acquisition process, volunteers were requested to fill out and sign consent forms. During the recording session, subjects were asked to perform seven different microexpressions defined by units in the Facial Action Coding System (FACS). FACS is a comprehensive framework that categorizes facial movements into distinct FAUs, with each FAU corresponding to a specific muscle movement in the face, such as raising the eyebrows or wrinkling the nose. These FAUs serve as the building blocks for identifying and analyzing facial expressions.

Table 2 lists the 27 Action Units (AUs) in the FACS. The seven microexpressions performed by the participants are combinations of FAUs as presented in Table 3: Smile (FAU 12), Brows Up (FAU 1), Nose Wrinkle (FAU 9), Open Mouth (FAU 25), One-Sided Lip Raise (FAU 10), Frown (a combination of FAUs 1, 2, and 4), and Chin Raise (FAU 17). These particular FAUs were chosen for their distinctiveness and their relevance to multiple emotions. By focusing on these FAUs, our dataset not only captures the physical movements but also allows for the exploration of how these movements correlate with different emotional states, providing a deeper understanding of microexpressions. Table 3 presents our proposed association between the selected microexpressions and the corresponding facial action units.

Each of the seven selected microexpressions was recorded six times per participant: three times under natural lighting and three times under studio lighting. This results in a balanced dataset, with approximately 120 videos per microexpression for each modality. For consistency and to avoid bias, we instructed participants only on the specific facial actions from the FACS codebook, without any reference to the underlying emotions these actions might represent. For example, they were asked to "raise eyebrows" without associating the action with emotions like fear or surprise. This approach ensured that the dataset captured pure facial movements rather than subjective emotional interpretations.

Table 2. Facial Action Units defined in the FACS. Each FAU is the result of a contraction or relaxation of one or more muscles.

AU	Name
1	Inn. brow raise
2	Out. brow raise
4	Brow lower
5	Upper lid raise
6	Cheek raise
7	Lower lid tight
9	Nose wrinkle
10	Lip Raise
11	Nasolabial
12	Lip corner pull
14	Dimpler
15	Lip corner depressor
16	Lower Lip depressor
17	Chin raise
18	Lip stretch
20	Lip tighten
23	Lip press
25	Lips part
26	Jaw drop
27	Mouth stretch

Table 3. Proposed microexpressions in the VETEX Face Dataset and their link to FACS FAUs.

Microexpression	FACS #	FACS Name
Smile	12	Lip corner puller
Brows up	1	Inner brow raiser
Nose wrinkle	9	Nose wrinkler
Open mouth	25	Lips part
One side lip raise	10	Upper lip raiser
Frown	1+2+4	Inner brow raiser Outer brow raiser Brow lowerer
Chin raise	17	Chin Raiser

For the synchronization of the two cameras, we employed a verbal instruction method during the recording session. Both cameras were set to record simultaneously, with the two experiment conductors initiating the recording at the same time after one of them gave a verbal instruction. Similarly, once the recording had started, one conductor would give a verbal cue to the participant, prompting them to perform the instructed facial action. Once the action was completed, the recording was stopped. During the recording process, the FLIR camera was connected via HDMI to a monitor, and the event camera was linked to a laptop running DV processing software, allowing real-time visualization of the recording. This setup ensured that the captured data met the research quality standards. If any issues were detected, such as misalignment or lighting inconsistencies, adjustments were made before proceeding to the next recording, and the affected sample was discarded.

3.3 Dataset Composition

The multi-modal VETEX dataset comprises a total of 2,506 videos of an average time of 4 s, distributed across three distinct modalities: 837 RGB videos, 828 thermal videos, and 841 event data videos. The final database comprises a total of approximately 2.75 h of video per modality. In the rare case where a recorded video from one modality was corrupted, only that specific video was discarded, while the corresponding videos in the other two modalities remained in the database. The recordings were collected from 20 participants, representing a diverse demographic group. The participant pool includes 15 males and 5 females, all within the age range of 20-30 years, and spans 10 different nationalities. To ensure the dataset's representativeness, 8 out of the 20 participants wore eyeglasses, introducing variability in facial appearance and potential occlusions, which further enriches the dataset.

The dataset is structured to facilitate comparisons between lighting conditions. For each participant and each microexpression, videos 1, 2, and 3 correspond to studio lighting, while videos 4, 5, and 6 correspond to natural lighting. To maintain a clear focus on facial movements, the dataset is annotated primarily with expression labels corresponding to the facial action units, without additional metadata. This approach emphasizes the raw facial dynamics, allowing for a pure analysis of microexpressions across the different modalities and lighting conditions. Example images from our dataset can be shown in Fig. 2.

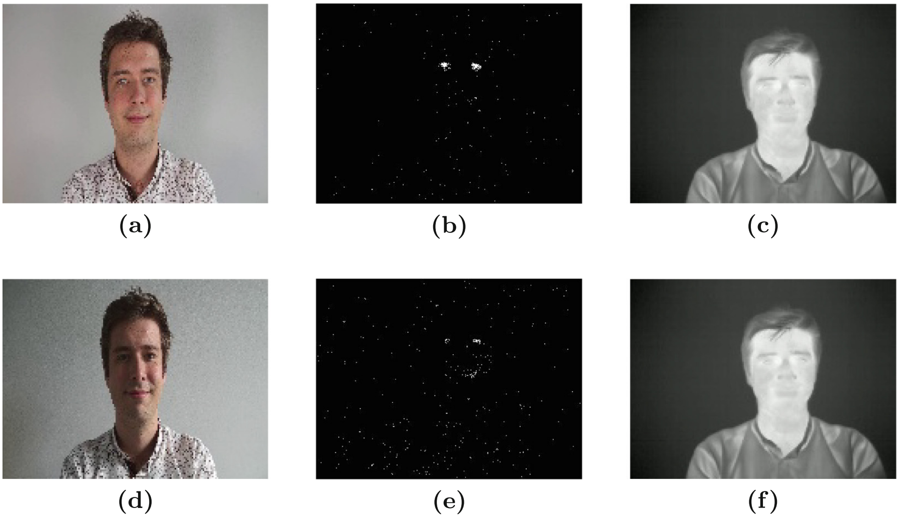

Fig. 2. Example frames from our VETEX dataset displayed in visible (left), event (center) and thermal (right) spectra. Frames (a-c) are recorded under studio light conditions; Frames (d–f) are recorded under natural light conditions.

4 Preliminary Assessment of the Dataset

In this section, we present the methodology followed in our work to compare the three spectra (RGB, thermal, and event) for the task of microexpression recognition. We also present our experimental results and evaluate the robustness of each modality under different lighting conditions.

4.1 Experimental Setup

Methodology. To assess the relevance of our dataset and evaluate the performance across these modalities, we conducted a series of experiments using a 3D CNN network, which was trained from scratch on video frames. The choice of a 3D CNN was driven by the nature of our dataset, which includes RGB and thermal videos, requiring a network capable of processing spatiotemporal information and capturing spatial patterns in video frames. We believe that this architecture choice delivers a good trade-off between a state-of-the-art network and a model capable of processing three different types of input data, to provide a fair comparison between the three spectra. Moreover, to incorporate event data into this network, we utilized the Temporal Binary Representation [13], which converts event streams into black-and-white frames, where a white pixel indicates at least one activated event within the frame in the selected time window. This representation is particularly effective for our purpose because it simplifies the event data into a format that highlights motion changes over time.