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 [21]. In addition, human faces reveal information about a person's emotional status, which has driven researchers to explore the possibility of automatically 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 [21]. 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, $3 - 8\mu m$) and long-wave infrared (LWIR, $8 - 15\mu m$) spectra [24], 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 hours 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: Section 2 presents advancements in the field of FER, and lists existing datasets containing thermal and event-based face videos. In Section 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, Section 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 [29]. 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 [28], 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 [17]. In another paper, Wang et al. [26] proposed a visible-thermal facial expression database and conducted experiments to analyze the relationship between facial temperature and emotion. More recently, Nguyen et al. [22] 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 [23], CK+[19], and JAFFE[20], 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 [21] 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 [26] containing 215 subjects, each displaying six expressions, and most recently, the KTFEv2 dataset [22] 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 [18], 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

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 [26]	-	215	✓	✓	×	✓	×	Emotions
2022	DFME [30]	10,045	97	✓	×	×	×	✓	Emotions
2018	TFAD [16]	2500*	90	×	✓	×	✓	✓	Landmarks/MicroExp
2022	DFME [30]	10,045	97	✓	×	×	×	✓	Emotions
2022	Becattini et al. [5]	455	25	✓	×	✓	×	×	Pos/Neg/Neutral
2023	LVT [21]	416	52	✓	✓	×	✓	×	Biometrics/eHealth
2023	NEFER [6]	609	29	✓	×	✓	×	✓	Emotions
2023	KTFEv2 [22]	1120	30	✓	✓	×	×	×	Emotions
2024	VETEX (Ours)	2506	30	✓	✓	✓	✓	×	MicroExp

FER-oriented datasets, we aimed to verify generalization under different lighting 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.

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 meters 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 seconds, 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 hours 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.

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.

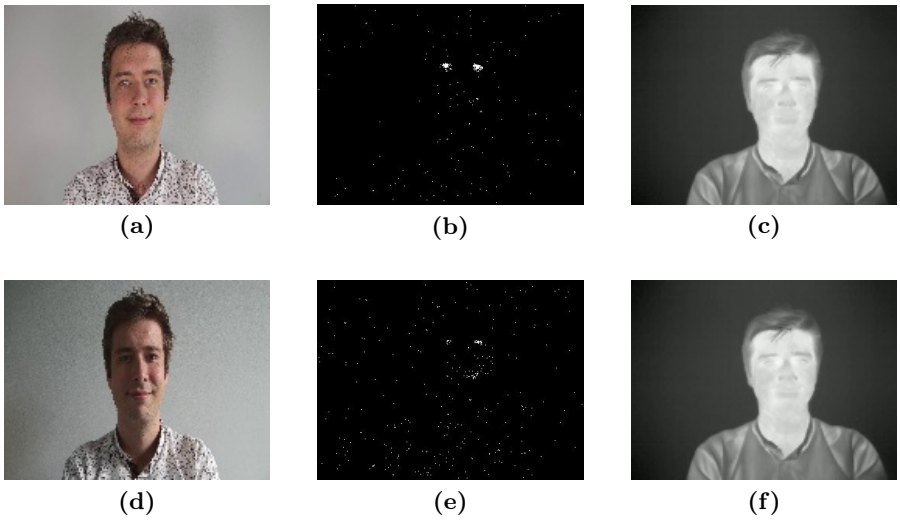

Fig. 2. Example frames from our VETEX dataset displayed in visible (left), event (center) and thermal (right) spectra. Frames (a-c) are recorded under studio light conditions; Frames (d-f) are recorded under natural light conditions.

4.1 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.

Implementation details: To ensure the robustness and reproducibility of our experiments, we carefully implemented the proposed methodology using consistent frameworks and parameters across all modalities.

The 3D CNN model was trained from scratch based on the implementation provided in [25], without any prior pre-training. The input frames were reshaped to a size of 116x116, and random flipping was applied to some of the frames.

For training, we performed a subject-exclusive split of the dataset to avoid any advantage due to data leakage. The training and test sets of VETEX consist of 14 and 6 people, respectively. To facilitate a fair training process and avoid bias by gender, we ensured an equal distribution of male and female volunteers in the train and test data splits. The 3D CNN was trained for 100 epochs with a learning rate of 0.001 and a batch size of 4.

For the event data preprocessing, we applied the EvFlow [27] denoising technique to remove any noise introduced by artificial lighting. Additionally, we transformed the event data into temporal binary frames using the same implementation as Innocenti et al. [13], with a window size of 15,000 and 8 bits. As for the thermal data the .TIFF files were converted into grayscale .AVI to be processed by the network. The RGB videos were processed in their initial .AVI format by the 3D CNN.

4.2 Results

Table 4 presents the performance of the 3D CNN for the two illumination conditions considered in the VETEX Face Dataset as well as the overall accuracy for each of the three data modalities. The results show that both event and thermal data outperform RGB in microexpression recognition, with event data improving performance by 12% and thermal data by 20% over RGB. Notably, thermal data achieved the highest overall accuracy at 34.81%. Moreover, thermal data consistently delivered the highest accuracy across both studio and natural lighting conditions. It is noteworthy that even without dedicated light sources, thermal data achieves state-of-the-art performance, likely due to its reliance on heat signatures rather than visible light. On the other hand, event data demonstrated a slight drop in accuracy under studio lighting conditions (26.98%) compared to natural lighting (27.55%), which might be due to overexposure noise introduced by artificial lighting. In the case of RGB data, an interesting behavior was observed, as the performance remained at 14.51% across both lighting conditions. This consistency could be attributed to the already low accuracy, which is similar to random classification when seven classes are present, as has been reported in other FER scenarios [6].

Overall, the results presented in this section highlight the significant advantage provided by the thermal and event modalities, even under varying environmental conditions.

5 Conclusion

This article introduces the multi-modal VETEX Face Database, containing approximately 8.24 hours of video from 20 different subjects under two different lighting conditions. The database comprises a total of 2,506 videos across two unconventional modalities in addition to RGB—thermal, and event—collected simultaneously using a paired visible-thermal camera (FLIR Duo R) and a DAVIS346 event camera. This dataset is designed to allow for the comparison or

Table 4. Evaluation of the baseline microexpression estimator on the VETEX test set for the three different input data modalities. Overall accuracy is reported, along with accuracy under two different lighting conditions: Studio and Natural. Accuracy is reported in % .

Modality	Total Accuracy	Studio Lighting Acc.	Natural Lighting Acc.
RGB	14.51%	14.51%	14.51%
Events	27.27%	26.98%	27.55%
Thermal	**34.81%**	**32.25%**	**35.77%**

fusion of the three data types in different facial processing tasks, such as FER. To the best of our knowledge, this is the first database providing visible-thermal-event face recordings, targeting subjects performing seven different microexpressions as defined by the facial action units in the Facial Action Coding System. The selected microexpressions can be associated with underlying emotions for FER, and they are not influenced by the subjective nature of feelings facilitating the task of deep learning models.

Experiments conducted on this novel dataset using a 3D CNN baseline demonstrate the superiority of thermal and event data over the visible spectrum. The images captured with the thermal camera deliver the best performance in both illumination conditions considered, as the thermal camera is able to specifically capture the subtle variations in heat signatures on the face, which can correspond to microexpressions. The event modality also proves successful in microexpression estimation, confirming that its high temporal resolution is better suited for capturing small facial movements than traditional RGB cameras. It is also proven that event data performs better in natural lighting conditions, where it captures details that might otherwise be missed in other domains. Building on these promising results, future work will explore the three modalities considered in the database in a fusion scenario, where the thermal and event data add different layers that complement the classic RGB spectrum.

Acknowledgments

This research is a part of the HEIMDALL project, funded by the BPI as part of the AAP I-Demo.

References

1. Iris database, accessed: Dec. 30, 2022. [Online]. Available: https://vciplokstate.org/pbvs/bench/Data/02/download.html
2. Nist equinox database, accessed: Dec. 30, 2022. [Online]. Available: http://www.equinoxsensors.com/products/HID.html

3. Adra, M., Dugelay, J.L.: Time-e2v: Overcoming limitations of e2vid. In: IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS). IEEE (2024)
4. Barchid, S., Allaert, B., Aissaoui, A., Mennesson, J., Djeraba, C.C.: Spiking-fer: spiking neural network for facial expression recognition with event cameras. In: Proceedings of the 20th International Conference on Content-based Multimedia Indexing. pp. 1–7. ACM (September 2023)
5. Becattini, F., Palai, F., Del Bimbo, A.: Understanding human reactions looking at facial microexpressions with an event camera. IEEE Transactions on Industrial Informatics **18**(12), 9112–9121 (2022)
6. Berlincioni, L., Cultrera, L., Albisani, C., Cresti, L., Leonardo, A., Picchioni, S., Becattini, F., Del Bimbo, A.: Neuromorphic event-based facial expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4109–4119 (2023)
7. Chen, J., Dey, S., Wang, L., Bi, N., Liu, P.: Multi-modal fusion enhanced model for driver's facial expression recognition. In: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–4. IEEE (2021)
8. Chowdary, M.K., Nguyen, T.N., Hemanth, D.J.: Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications **35**(32), 23311–23328 (2023)
9. Corneanu, C.A., Simón, M.O., Cohn, J.F., Guerrero, S.E.: Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. IEEE transactions on pattern analysis and machine intelligence **38**(8), 1548–1568 (2016)
10. Eddine, M.J., Dugelay, J.L.: Gait3: An event-based, visible and thermal database for gait recognition. In: 2022 International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–5. IEEE (2022)
11. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(1), 154–180 (2020)
12. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1312–1321 (2021)
13. Innocenti, S.U., Becattini, F., Pernici, F., Del Bimbo, A.: Temporal binary representation for event-based action recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 10426–10432. IEEE (2021)
14. Karnati, M., Seal, A., Bhattacharjee, D., Yazidi, A., Krejcar, O.: Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. IEEE Transactions on Instrumentation and Measurement **72**, 1–31 (2023)
15. Kopaczka, M., Kolk, R., Merhof, D.: A fully annotated thermal face database and its application for thermal facial expression recognition. In: 2018 IEEE international instrumentation and measurement technology conference (I2MTC). pp. 1–6. IEEE (2018)
16. Kopaczka, M., Kolk, R., Merhof, D.: A fully annotated thermal face database and its application for thermal facial expression recognition. In: 2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). pp. 1–6. IEEE (2018). https://doi.org/https://doi.org/10.1109/i2mtc.2018.8409768

17. Lee, J., Kim, S., Kim, S., Sohn, K.: Multi-modal recurrent attention networks for facial expression recognition. IEEE Transactions on Image Processing **29**, 6977–6991 (2020)
18. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Transactions on Affective Computing **13**(3), 1195–1215 (2022). https://doi.org/https://doi.org/10.1109/taffc.2020.2981446
19. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 ieee computer society conference on computer vision and pattern recognition-workshops. pp. 94–101. IEEE (2010)
20. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Proceedings Third IEEE international conference on automatic face and gesture recognition. pp. 200–205. IEEE (1998)
21. Mirabet-Herranz, N., Dugelay, J.L.: Lvt face database: A benchmark database for visible and hidden face biometrics. In: 2023 International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–6. IEEE (2023)
22. Nguyen, H., Tran, N., Nguyen, H.D., Nguyen, L., Kotani, K.: Ktfev2: multimodal facial emotion database and its analysis. IEEE Access **11**, 17811–17822 (2023)
23. Qu, F., Wang, S.J., Yan, W.J., Fu, X.: Cas (me) 2: A database of spontaneous macro-expressions and micro-expressions. In: Human-Computer Interaction. Novel User Experiences: 18th International Conference, HCI International 2016, Toronto, ON, Canada, July 17-22, 2016. Proceedings, Part III 18. pp. 48–59. Springer (2016)
24. Rai, M., Maity, T., Yadav, R.: Thermal imaging system and its real time applications: a survey. Journal of Engineering Technology **6**(2), 290–303 (2017)
25. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). IEEE (2015). https://doi.org/https://doi.org/10.1109/iccv.2015.510
26. Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Chen, F., Wang, X.: A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Transactions on Multimedia **12**(7), 682–691 (Nov 2010)
27. Wang, Y., Du, B., Shen, Y., Wu, K., Zhao, G., Sun, J., Wen, H.: Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6358–6367 (2019)
28. Zhang, X., Yang, X., Zhang, W., Li, G., Yu, H.: Crowd emotion evaluation based on fuzzy inference of arousal and valence. Neurocomputing **445**, 194–205 (2021)
29. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image and vision computing **29**(9), 607–619 (2011)
30. Zhao, S., Tang, H., Mao, X., Liu, S., Zhang, Y., Wang, H., Xu, T., Chen, E.: Dfme: A new benchmark for dynamic facial micro-expression recognition. IEEE Transactions on Affective Computing (2023)