Beyond the Smile: Capturing Micro Expressions with Event-based Vision

semester project

Victor MAYAUD[1], Elliot BOUCHY[1]

[1] Data Science, EURECOM, France

Abstract

Micro-expressions, the brief and involuntary facial expressions that reveal true emotions, are pivotal in understanding human affective states. Traditional methods of capturing these fleeting expressions often fail due to their limited temporal resolution and high latency. This study explores the efficacy of event-based vision systems in detecting and analyzing micro-expressions. Unlike conventional frame-based cameras, event-based sensors capture changes in the visual scene asynchronously with high temporal precision, providing a robust solution for recording rapid facial movements. We developed an innovative framework integrating event-based vision with advanced machine learning algorithms to detect and classify micro-expressions. Experimental results demonstrate significant improvements in both accuracy and processing speed compared to traditional methods. This approach opens new avenues for applications in psychology, security, and human-computer interaction, offering a deeper insight into human emotions beyond the visible smile.

1 Introduction

In the rapidly evolving landscape of human-computer interaction, understanding and accurately interpreting human emotions have become increasingly essential. Traditional methods of emotion detection often rely on facial expressions, yet they are limited by the capabilities of standard cameras, which may miss subtle cues and rapid changes in facial dynamics. Micro-expressions, fleeting and involuntary facial movements lasting only a fraction of a second, are particularly challenging to capture using conventional RGB cameras. These micro-expressions provide crucial insights into an individual's emotional state, revealing underlying feelings that may not be consciously expressed.

To address this limitation, this project delves into the innovative realm of event-based vision for Facial Expression Recognition (FER). Event cameras, inspired by the human visual system, operate differently from conventional cameras by responding to changes in the scene rather than capturing frames at fixed intervals. This motion-driven approach enables event cameras to capture data only when there are changes in pixel intensities, resulting in a sparse yet high-resolution representation of dynamic scenes.

The utilization of event cameras offers a paradigm shift in FER, enabling the detection of micro-expressions with unparalleled accuracy and temporal resolution. By leveraging the unique capabilities of event cameras, this project aims to develop advanced algorithms capable of precisely recognizing and interpreting a broad spectrum of human emotions, including those expressed through subtle micro-expressions.

Beyond academic curiosity, the implications of this research extend to practical applications in various domains. Improved emotion recognition facilitates more natural and intuitive human-machine interaction, enhancing user experience and communication in fields such as virtual reality, human-computer interfaces, and affective computing. Moreover, the ability to capture micro-expressions in challenging environments, such as low-light conditions or fast-paced dynamic settings, opens avenues for applications in security, healthcare, and behavioral analysis.

This project offers an exciting opportunity to engage in cutting-edge research at the intersection of computer vision, artificial intelligence, and psychology.

2 Emotion

2.1 Context

An emotion is a complex reaction of the organism that manifests itself through physiological, behavioral and cognitive responses to an internal or external stimulus. Emotions play a central role in the way human beings interact with the world around them, influencing decisions, actions and interpersonal relationships. They are often classified into several basic categories, such as joy, fear, sadness, disgust, anger and surprise.

- **Joy:** Joy is a positive emotion often expressed through distinct facial movements. Expressions associated with joy include opening the mouth and smiling, sometimes accompanied by dimples. When someone is extremely happy, they may open their mouth to laugh, exposing their teeth and raising the corners of their mouth. A genuine smile, also known as a Duchenne smile, engages the muscles around the eyes, causing wrinkles at the corners of the eyes and dimples on the cheeks.
  The detection of joy using event-driven vision systems relies on the ability of these sensors to rapidly register subtle movements of the lips and eyes. For example, when the corners of the lips lift, the sensor captures these changes almost instantaneously, enabling precise analysis of the smile. This technology overcomes the limitations of traditional cameras in terms of temporal resolution, offering a detailed capture of the micro-expressions associated with joy.
- **Fear:** Fear is an intense emotion often triggered by a perception of threat or danger. Typical facial expressions of fear include mouth opening and frowning. Opening the mouth may be a reaction of surprise or preparation to scream, while frowning indicates concentration or an attempt to understand the source of the fear.
  Event-driven vision sensors can capture these micro-expressions with remarkable precision. When a person rapidly opens their mouth and frowns, these sensors detect the changes in real time, recording each movement in detail. This ability to capture rapid, synchronized movements is essential for accurate analysis of fearful expressions, which can otherwise go unnoticed.
- **Sadness:** Sadness is often expressed by micro-expressions such as frowning and blinking. Frowning in a state of sad-

ness creates an appearance of dismay or grief, while blinking may indicate a struggle to hold back tears or an effort to mask sadness.
Event-driven vision systems can capture these subtle expressions with unprecedented precision. The blink, in particular, is a micro-expression that occurs in a fraction of a second and can be easily missed by traditional capture methods. Event sensors record these rapid movements, enabling detailed analysis of the signs of sadness on a person's face.

- **Disgust:** Disgust is a negative emotional reaction often triggered by unpleasant or repulsive stimuli. Typical facial expressions of disgust include an inverted smile (where the corners of the lips are turned downwards), a wrinkle of the nose and a frown. These micro-expressions reflect intense aversion or rejection.
  The detection of disgust via event-driven vision systems relies on the simultaneous capture of several distinct micro-expressions. Sensors record changes in the muscles around the mouth, nose and eyebrows, providing a complete view of the disgusted expression. The ability of these sensors to capture rapid, synchronized movements is crucial for accurate analysis of this complex emotion.
- **Anger:** Anger often manifests itself through the contraction of the masseter muscles (jaw muscles) and a pronounced frowning of the eyebrows. These micro-expressions indicate the tension and aggression characteristic of anger. Masseter contraction is particularly indicative of intense anger, as it prepares the body for a possible physical reaction.
  Event-driven vision systems capture these rapid, intense movements with great precision. Contraction of the masseter muscles and frowning often occur simultaneously, and sensors can record these changes in real time. This ability to detect complex micro-expressions enables a deeper and more precise understanding of anger states.
- **Surprise:** Surprise is an emotion often triggered by unexpected events. Typical facial expressions of surprise include opening the mouth and rapidly raising the upper eyelids. These micro-expressions indicate an immediate, involuntary reaction to something unexpected.
  Event-driven vision systems are particularly effective at capturing expressions of surprise, thanks to their ability to record rapid, sudden changes. When the mouth opens and the eyelids lift, these sensors record these movements almost instantaneously. This ability to detect rapid, synchronized micro-expressions enables precise analysis of surprise reactions.

2.2 Micro-expression

Micro-expressions are very rapid, involuntary facial movements that occur in response to emotions experienced by a person. These expressions, which generally last between $1/25^{th}$ and $1/5^{th}$ of a second, are often too rapid to be consciously perceived by the human eye. Unlike longer-lasting, voluntary facial expressions, micro-expressions occur spontaneously and uncontrollably, revealing genuine emotions that the individual may be trying to conceal. Because of their brevity and subtlety, micro-expressions offer an authentic insight into internal emotional states, and are of great interest in fields such as psychology, lie detection and social interaction. The ability to identify and analyze these micro-expressions can provide valuable information about underlying emotions, often hidden by conscious, controlled facial expressions.[8]

Figure 1. Examples of Facial Expressions and Corresponding Action Units

2.3 Detection of the emotions

Our research focuses on the use of event-driven vision systems to identify human emotions through the precise detection of micro-expressions. Thanks to this technology, we have been able to record extremely rapid and subtle facial movements, often undetectable by conventional cameras. Here's how we identified six specific emotions by analyzing combinations of micro-expressions

- **Joy** To detect joy, we targeted two main micro-expressions: the opening of the mouth and the smile. Event-driven vision systems are particularly effective at capturing lip corner elevation and dimple formation, key indicators of a genuine smile. In addition, the opening of the mouth, often associated with laughter, is rapidly detected thanks to the high temporal resolution of the sensors. These sensors record changes in the muscles around the mouth in real time, enabling precise identification of joy.
- **Fear** Fear often manifests itself as a combination of an open mouth and a frown. Event sensors simultaneously capture these two micro-expressions by detecting rapid changes in the facial muscles. The sudden opening of the mouth, often indicating a cry or a reaction of surprise, is coupled with a frown, a sign of tension and concentration. This ability to record simultaneous movements enables precise identification of expressions of fear.
- **Sadness** Sadness is often indicated by frowning and frequent blinking. Event-driven vision systems capture these micro-expressions by recording rapid eyelid movements and forehead muscle contractions. Blinking, which may be an attempt to hold back tears, is detected in real time, while frowning, indicating grief or distress, is also accurately captured.
- **Disgust** Disgust is manifested by a complex combination of micro-expressions, including an inverted smile (where the corners of the lips are directed downwards), a wrinkle of the nose and a frown. Event sensors are able to detect these subtle and multiple changes simultaneously. The technology records muscle movements around the mouth and nose, as well as forehead contractions, providing a detailed analysis of this negative emotion.
- **Anger** Anger is often visible through the contraction of the jaw muscles and a pronounced frown. Event-driven vision systems capture these micro-expressions by detecting rapid, intense contractions of the jaw and forehead muscles. The precision of these sensors makes it possible to

identify even the most subtle forms of anger, by recording the synchronized movements of facial muscles.

- **Surprise** Surprise is characterized by an open mouth and rapid elevation of the upper eyelids. Event sensors are particularly effective at capturing these sudden changes. When a person is surprised, their mouth opens wide and their eyelids rise, movements detected almost instantaneously by the sensors. This ability to record rapid, coordinated facial expressions enables precise identification of surprise.

3 Event-camera

3.1 Introduction

Event-driven cameras represent a major advance in image capture technology, offering significant advantages over traditional cameras. Unlike conventional image sensors, which record scenes by taking still images at regular intervals (i.e., frames per second), event-driven cameras work by capturing only changes in the visual scene. This approach enables extremely high temporal resolution, essential for detecting rapid, subtle movements such as facial micro-expressions.[2]

3.2 The functioning of event cameras

Event cameras use a fundamentally different approach to image capture than traditional cameras. Here's a detailed look at how they work:

Independent and Asynchronous Pixels: Each pixel in the event-camera operates independently and reacts asynchronously to changes in light intensity. This means that each pixel can detect and report a change without waiting for a full image capture cycle.

Event detection : An "event" is generated when the change in light intensity at a pixel exceeds a predetermined threshold. This threshold can be adjusted to sensitize the camera to specific variations in light intensity.

Event data : Each event comprises three main pieces of information:

- **Spatial coordinates (x, y):** The position of the pixel where the change was detected.
- **Timestamp:** The exact moment when the event took place, usually measured in microseconds.
- **Polarity:** The direction of the change in intensity (increase or decrease).

Transmission and processing: Events are transmitted and processed in real time. The data flow generated is much lower than that of traditional cameras, as only changes are transmitted, reducing the bandwidth required.

3.3 Advantages of Event Cameras

Event-driven cameras offer several distinct advantages over traditional cameras, particularly in applications requiring fast, accurate motion capture.

High temporal resolution:

Event-driven cameras can capture motion with microsecond temporal resolution, far surpassing traditional cameras. This capability is essential for analyzing very fast-moving phenomena, such as facial micro-expressions or rapid mechanical movements.

Energy efficiency:

By recording only changes in the scene, event cameras consume less energy. The sensors do not require constant lighting, thus reducing energy consumption.

This efficiency is particularly beneficial for real-time applications and embedded devices requiring low power consumption.

Low latency:

The asynchronous nature of event capture minimizes latency, as information is transmitted instantaneously as soon as a change is detected.

This enables near real-time reaction, essential for applications such as robotics, autonomous vehicles and emotion detection, where rapid responses are critical.

Robustness to Variable Light Conditions:

Event cameras operate effectively in low-light or variable-light environments. Because they respond to changes in light rather than absolute light levels, they can capture events in conditions where traditional cameras would struggle.

This makes them ideal for surveillance and security applications, where lighting conditions can be unpredictable.

3.4 Comparison with traditional RGB cameras

Traditional RGB cameras capture images using a matrix sensor that records color information (red, green, blue) for each pixel at fixed intervals. Here's a detailed comparison of the two technologies:

Capture method :

RGB Camera: Captures images at fixed frame rates (e.g. 30 frames per second), where each frame contains complete information about the scene.

Event Camera: Captures only changes in the scene, with each pixel operating independently and asynchronously.

Temporal resolution :

RGB camera: Temporal resolution is limited by the frame rate. For example, a 30 fps camera has a temporal resolution of 33 ms.

Event Camera: Temporal resolution can be on the order of microseconds, enabling extremely fast events to be captured.

Data efficiency :

RGB Camera: Generates a large amount of data, as each image contains complete information on the scene, regardless of movement or change.

Event Camera: Generates less data by capturing only changes, reducing bandwidth requirements and improving processing efficiency.

Lighting conditions:

RGB Camera: Can have difficulties in low light or changing light conditions, often requiring adjustments to gain or exposure time.

Event Camera: Works effectively in low-light conditions, as it detects changes in light rather than absolute light levels.

Applications :

RGB Camera: Used in applications where full image capture is required, such as photography, video and some forms of surveillance.

Event Camera: Ideal for applications requiring fast, accurate motion capture, such as robotics, autonomous vehicles, micro-expression detection and real-time security systems.

Event cameras offer a revolutionary technology for image capture, surpassing traditional RGB cameras in terms of temporal resolution, energy efficiency and robustness to varying light conditions. Their ability to detect rapid, subtle movements opens up new possibilities in many fields, including robotics, autonomous vehicles, security surveillance and emotional analysis. By capturing only changes in the scene, they enable more precise and efficient analysis, particularly suited to real-time applications where speed and precision are crucial.

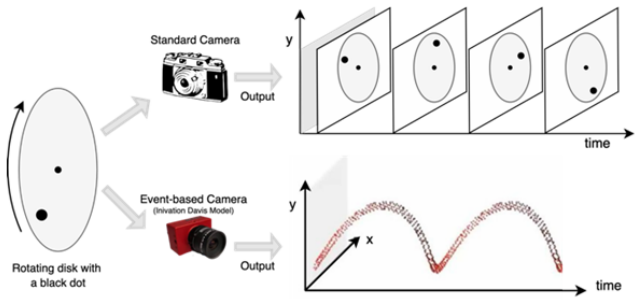

Figure 2. event camera versus RGB camera.

4 Spiking Neural Network

4.1 context

Spiking Neural Networks (SNNs) are a class of artificial neural networks (ANNs) that more closely mimic the behavior of natural neural networks. Unlike typical multi-layer perceptron networks, which transmit information at every propagation cycle, SNNs integrate the concept of time into their operational model. Neurons in an SNN transmit information only when their membrane potential—the intrinsic electrical charge across the neuron's membrane—reaches a specific threshold. At this point, the neuron fires, generating a signal that propagates to other neurons, which then adjust their membrane potentials in response to this signal. This model of neuron, which fires upon threshold crossing, is referred to as a spiking neuron model.

Detecting micro-expressions is well-suited to SNNs due to their dynamic and temporally sensitive nature. For our project, we employed an event camera that captures facial movements by detecting changes in pixel intensity, analogous to the way neurons respond to stimuli. An event camera does not record static frames but rather detects changes in the scene, triggering an output only when a change in intensity is detected—akin to the firing of neurons. These intensity variations are processed as inputs to the SNN, where a combination of such pulsations from changed pixel intensities might collectively reach the threshold, thus triggering a neuronal response. This makes SNNs particularly effective for analyzing the high-speed, subtle facial movements characteristic of micro-expressions.

4.2 Integrate-And-Fire-Models

The Integrate-and-Fire model is a fundamental neuron model used in spiking neural networks (SNNs). It is designed to simulate the behavior of biological neurons, which integrate incoming electrical signals until a threshold is reached, triggering a spike or action potential. This model captures the essential characteristics of neuronal activity without the complexity of biophysically detailed models.

4.2.1 Mathematical Formulation The dynamics of the membrane potential in an Integrate-and-Fire neuron are described by the following differential equation:

$$\tau \frac{du}{dt} = -(u - u_{\text{rest}}) + RI(t)$$

where $u(t)$ represents the membrane potential at time t, τ is the membrane time constant, u_{rest} is the resting membrane potential, R is the membrane resistance, and $I(t)$ is the input current.

When the membrane potential $u(t)$ reaches a certain threshold θ, the neuron fires a spike, and the membrane potential is subsequently reset to a lower value, typically below u_{rest}. After firing, the neuron enters a refractory period during which it cannot spike again, regardless of the input.

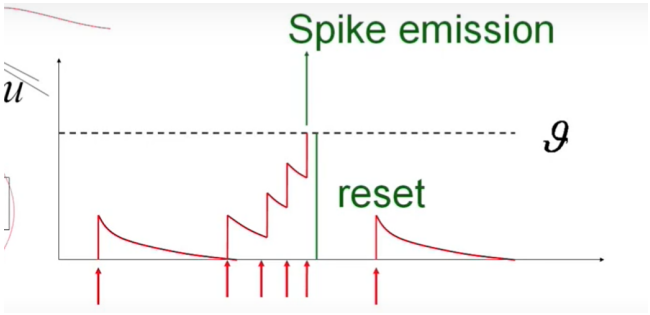

Figure 3. Diagram illustrating the spike emission and reset mechanism in an Integrate-and-Fire neuron.

4.2.2 Spike Emission and Reset Mechanism As shown in the figure, the process of spiking and resetting is crucial for temporal coding in neural networks. It allows neurons to encode information in the timing of spikes rather than the rate of firing, which is more akin to how biological neural systems operate.

This model is particularly effective in applications like micro-expression detection using event cameras, where temporal precision is crucial. The event-driven nature of both the camera and the SNN ensures that the system is highly responsive and energy-efficient, capturing subtle facial movements in real-time.

4.3 2D Leaky Integrate-and-Fire

The 2D leaky integrate-and-fire model is an extension of the basic integrate-and-fire model, incorporating an additional dimension to the neuron's dynamic properties. This model is designed to better capture the complexities of neuronal behavior by allowing the threshold for firing to vary based on the neuron's recent activity.

4.3.1 Model Description In this model, each neuron's membrane potential V is updated based on the incoming spikes and its current state. Unlike the standard model, which has a fixed threshold, the 2D model allows the threshold V_t to adapt over time. This adaptation is critical in environments where the stimulus characteristics are highly variable, as it allows the neuron to modulate its sensitivity dynamically.

4.3.2 Dynamics After a Spike After the neuron fires a spike, the membrane potential is reset to zero, and the threshold V_t is increased by a small increment δV_t. This mechanism is depicted in the following equations:

$$V \leftarrow 0$$

$$V_t \leftarrow V_t + \delta V_t$$

The adaptation of the threshold is intended to prevent the neuron from firing too frequently within a short time, a phenomenon known as the refractory period. This adaptation can be crucial for processing signals that require fine temporal resolution, such as the detection of micro-expressions in facial recognition systems. We can see an illustration of 2D leaky integrate-and-fire model in the figure bellow 4

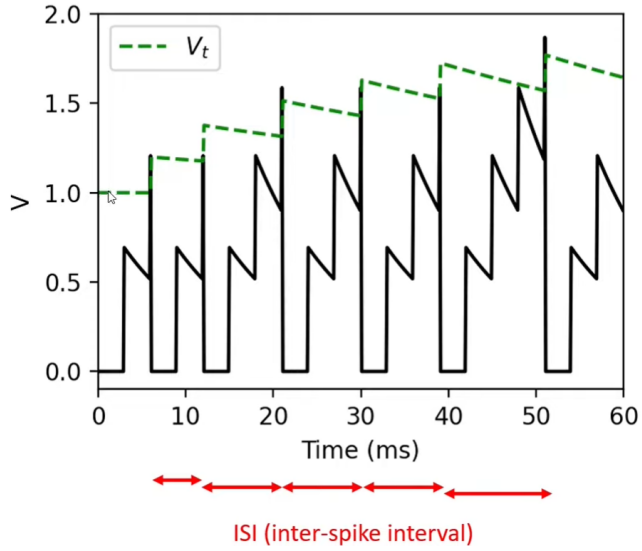

Figure 4. Graph showing the adaptation of the threshold V_t and the membrane potential reset in the 2D leaky integrate-and-fire model.

4.3.3 Impact on Neural Coding The dynamic threshold introduces a form of plasticity into the neuron's response, allowing it to be more selective about when to fire based on both the current and past inputs. This feature is particularly useful in spiking neural networks used for tasks that require discrimination of subtle and rapid changes in the input, such as those captured by event cameras in micro-expression detection.

By adapting the firing threshold over time, the 2D leaky integrate-and-fire model provides a more robust mechanism for handling varying input intensities, thereby enhancing the network's ability to detect and respond to significant events within noisy environments.

4.4 Structure of the neural network

In order to construct the spiking neural network (SNN), we utilize the library SpikingJelly, which is an extension of PyTorch tailored for SNN applications. This library retains many of PyTorch's functionalities but replaces traditional neurons with spiking neurons, accommodating the temporal dynamics essential for processing data from event-based sensors.

The architecture of our SNN closely resembles that of a convolutional neural network (CNN), but with significant modifications to leverage the characteristics of spiking neurons. Following the initial convolutional and pooling layers, which are crucial for extracting and downsampling features from the input video, spiking neuron layers are integrated. These layers focus on processing dynamic changes and detecting crucial characteristics from the event-driven video data. This layered approach ensures that each stage of the network adds to the selectivity and invariance of the features, enhancing the detection of subtle facial movements crucial for interpreting micro-expressions.

The input to our network consists of video data from an event camera, formatted as numpy arrays (.npy files). This format is particularly efficient for handling large datasets and facilitates rapid processing within the neural network. The output of the network is the detected facial movements, which are identified based on the patterns of spikes generated by the final spiking neuron layers.

For this project, we employ a pre-designed spiking neural network model available on GitHub under the repository named 'Spike-Element-Wise-ResNet'. This model provides a robust framework that we can further customize and adapt through modifications in the code. By adjusting parameters and integrating new layers, we can optimize the network to better respond to the specific nuances of facial expressions captured by the event camera.

By utilizing 'Spike-Element-Wise-ResNet', we leverage a well-established model that has been proven effective in other applications, ensuring our foundation is both solid and capable of being tailored to the specific challenges and requirements of detecting micro-expressions in real-time video data.

5 Dataset creation

To train our spiking neural network, we require a substantial dataset of diverse facial movements. Given the scarcity of suitable resources online, particularly videos captured with event cameras, we have opted to compile our own dataset. Utilizing our event camera, we aim to create a robust and usable dataset that specifically addresses the unique requirements of our project.

5.1 Online Resources and Datasets

In the development of our spiking neural network, we reviewed several online datasets that are commonly used for facial expression and micro-expression detection. These datasets vary significantly in their focus and format, influencing their suitability for different types of neural network architectures.

- **NEFER:** This dataset is primarily designed for convolutional neural networks (CNNs) and includes categorized emotions. It is not specifically optimized for SNNs, which may limit its direct applicability to our project [4].
- **FES (Faces in Event Stream):** Featuring 689 minutes of event camera data, this dataset is exclusively geared towards face recognition tasks rather than dynamic facial movement or expression detection [5].
- **CASME II (Chinese Academy of Sciences Micro-expression):** This dataset is tailored for micro-expression analysis with videos recorded at 200 frames per second and a resolution of 280x340 pixels. It categorizes expressions into 5 classes, providing a nuanced view of facial expressions [6].
- **Nexdata/57 Types of Micro-expression Data:** This extensive dataset includes micro-expression video data from over 2,000 individuals across different ethnicities (Asian, Black, Caucasian, and Brown). It is one of the most diverse collections available, supporting a wide range of facial recognition research [7].

While these datasets provide a valuable foundation for facial expression analysis, they often exhibit significant imbalances in the distribution of categories. Additionally, most are not specifically designed for detecting subtle facial movements, which is critical for micro-expression recognition. This highlights the necessity for creating a custom dataset using our event camera, which can more precisely capture the high-speed and subtle facial dynamics integral to our research on micro-expressions.

5.2 protocol

The protocol for recording micro-expressions was meticulously crafted to ensure accurate capture and analysis of subtle facial dynamics using an event camera and a thermal camera. This section outlines the setup and procedure adopted for the dataset creation.

- Subjects were positioned in front of a white background to avoid visual distractions and interference. A specific distance of 30-40 cm was maintained between the subjects and both cameras to ensure optimal focus and frame composition. The setup aimed to standardize the recording environment across all sessions.
- Lighting was carefully adjusted to minimize shadows and uniformly highlight facial features. The primary goal was to ensure even illumination across the subject's face, enhancing the visibility of micro-expressions and facilitating their accurate detection by the cameras.
- Each subject was instructed to maintain a neutral facial expression initially. A specific micro-expression was assigned to each subject prior to recording. Subjects were trained to perform the micro-expression on cue, ensuring consistency and reliability in the responses captured.
- To synchronize the event camera with the thermal camera, a piece of paper was placed in front of both cameras. This not only served to mark the start of the video recording but also ensured that both cameras were perfectly aligned in terms of timing, crucial for simultaneous data capture.
- Recording commenced simultaneously on both cameras after the removal of the paper, following a brief two-second countdown. This method guaranteed that the recording phase started at exactly the same moment for both devices, capturing the entire sequence of facial expressions.
- Five seconds after the recording started, the subject was signaled to perform the designated micro-expression. This cue was typically given by a simple hand gesture, ensuring that the subject's performance was timed accurately relative to the recording.
- The recording session was concluded after ten seconds. This duration was chosen to balance the need for capturing a complete expression while keeping the recording brief enough to maintain the subject's comfort and the session's efficiency.

This recording protocol was designed to optimize the capture of micro-expressions using both an event camera and a thermal camera, enabling detailed analysis of these rapid, subtle facial movements.

5.3 Dataset Description

The dataset made for this analysis encompasses a total of 233 videos. Each video has an average duration of approximately 8 seconds, resulting in a cumulative total of around 30 minutes of video footage. This dataset is rich in diversity, capturing various facial expressions which are essential for comprehensive facial expression recognition studies.

The facial expressions are categorized into the following distinct classes :

- **Upper Lid Raiser**: 14.2%
- **Smile**: 13.3%
- **Open Mouth**: 17.2%
- **Nose Wrinkle**: 13.3%
- **Frown**: 15.0%
- **Blink Eyes**: 13.7%
- **Contract Jaw**: 13.3%

As depicted in Figure 5, the pie chart illustrates the proportional representation of each facial expression category within the dataset. The "Open Mouth" expression is the most prevalent, constituting 17.2% of the dataset, while "Smile" and "Nose Wrinkle" both represent 13.3%. The other categories, "Upper Lid Raiser," "Frown," "Blink Eyes," and "Contract Jaw," are distributed relatively evenly, each comprising around 13-15% of the dataset.

This dataset provides a balanced and comprehensive foundation for analyzing facial expressions, allowing for detailed examination and training of facial recognition models. The diversity in expression types ensures that the models trained on this data can generalize well to a wide range of facial movements and expressions.

In summary, the dataset's extensive coverage of various facial expressions, along with its substantial video duration, makes it a valuable resource for research and development in the field of facial expression recognition. The balanced distribution of expression categories further enhances its utility for training robust and accurate recognition models.

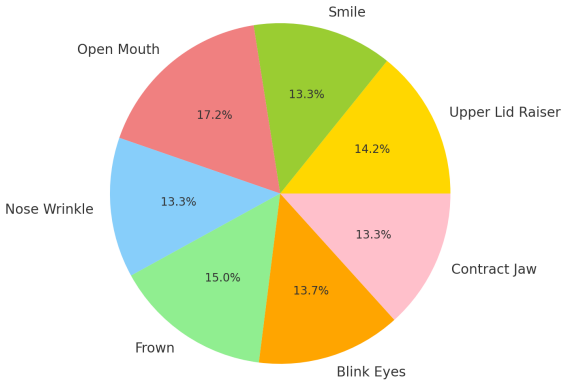

Figure 5. Distribution of Facial Expressions in the Dataset

6 Results

6.1 Accuracy on Seven Categories

The graph labeled "test_acc1" 6 displays the accuracy of the SEWResNet model in classifying data into seven distinct categories. Initially, the accuracy shows a gradual increase from approximately 25% at the start to around 35% after 20 epochs. This upward trend continues with some fluctuations, indicating the model's learning process as it adapts to the data's complexity. After 66 epochs, the accuracy stabilizes, reaching a peak precision of 43.18%. The smoothed accuracy value at this point is approximately 40.24%, suggesting that the model has effectively learned to categorize the data but still has room for improvement.

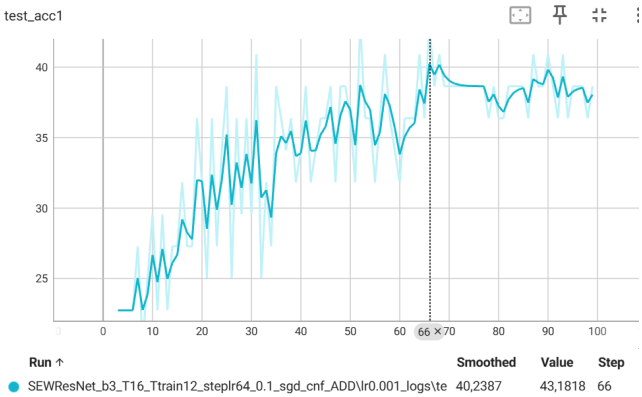

Figure 6. Accuracy of the model SEWResNet

6.2 Accuracy on Specific Subset

The second graph, "test_acc5," 7 represents the model's accuracy on a specific subset of the dataset. Here, the initial accuracy is significantly higher, starting at around 85%, which suggests that the model finds this subset less challenging. The accuracy reaches its maximum value of 95.45% at epoch 20, reflecting the model's strong performance on this particular data segment. However, after this peak, the accuracy exhibits a progressive decline, stabilizing around 86%. This pattern might indicate overfitting to the initial data followed by adjustments as the model encounters more diverse examples within the subset. The final smoothed accuracy recorded is 92.43%, showing a consistently high performance overall.

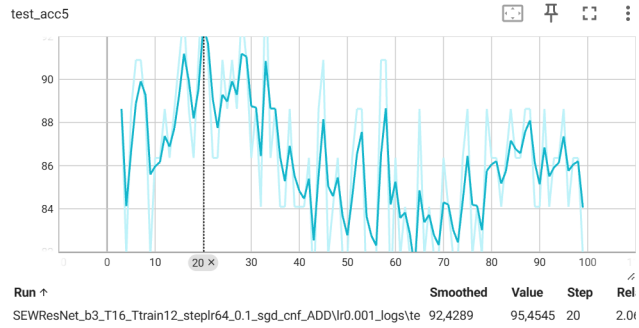

Figure 7. Accuracy Trend Over Epochs for Test Accuracy 5

6.3 Loss Function Analysis

The "test_loss" 8 graph provides insights into the model's loss values during training, using the cross-entropy loss function. Initially, the loss starts high at approximately 2.4, indicating the model's struggle to correctly classify the data. Over the first 20 epochs, the loss decreases steadily, reaching a minimum value of about 1.8 around epoch 22. This reduction in loss corresponds with the initial improvements in accuracy observed in the other graphs. However, after reaching this minimum, the loss begins to increase again, suggesting potential issues such as overfitting or encountering more difficult data instances as training progresses. By epoch 99, the loss stabilizes around 2.39, highlighting the need for further refinement to maintain lower loss values consistently.

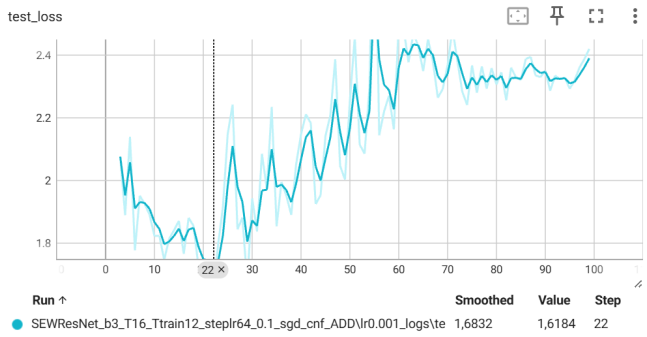

Figure 8. Cross entropy of the model SEWResNet

6.4 Comparison

In comparison to the work conducted by Berniconi [1] in the research paper titled "Neuromorphic Event-based Facial Expression Recognition," our results demonstrate significant improvements in emotion detection accuracy. Berniconi utilized a Convolutional Neural Network (CNN) to detect emotions from the NEFER dataset, achieving an accuracy of 30.95%.

Our approach, leveraging a Spiking Neural Network (SNN) combined with event-based vision, surpasses this performance. As shown in the accuracy trends, our model achieves a peak accuracy of 43.18% on the seven-category classification task and stabilizes at approximately 40.24%. Furthermore, on a specific subset of the dataset, our model reaches a maximum accuracy of 95.45% and stabilizes around 86%, showcasing its robustness and superior performance.

These results highlight the efficacy of spiking neural networks in handling the rapid, dynamic nature of micro-expressions, providing a more accurate and reliable method for facial emotion detection compared to traditional CNN approaches as demonstrated by Berniconi.

7 Discussion

For this project, the event-based vision system demonstrated its effectiveness in capturing and analyzing micro-expressions, achieving significant improvements in both accuracy and processing speed compared to traditional frame-based cameras. The high temporal resolution and low latency of event cameras enabled the detection of rapid and subtle facial movements, providing a more accurate analysis of emotions.

7.1 Effectiveness of the Event-based Vision System

The event-based vision system showed superior performance in detecting micro-expressions. The system's high temporal resolution allowed for the precise capture of quick facial movements, which are often missed by conventional cameras. This capability is particularly beneficial in applications where rapid response times are critical, such as security and human-computer interaction. The integration of advanced machine learning algorithms further enhanced the system's ability to classify micro-expressions accurately, demonstrating a significant improvement over traditional methods.

7.2 Limitations of the Approach

Despite its advantages, the event-based vision system has some limitations. One major challenge is the need for sophisticated algorithms to process the sparse and asynchronous data generated by event cameras. Developing these algorithms requires significant computational resources and expertise in machine learning. Additionally, event cameras can be sensitive to noise, which may affect the accuracy of emotion detection in certain environments. Ensuring consistent performance across varying lighting conditions and facial expressions remains an ongoing challenge.

7.3 Future Work

To further improve the performance of micro-expression detection systems, several research avenues can be explored.

We could explore more advanced learning techniques, such as reinforcement learning or transfer learning, which could improve the model's ability to generalize. These methods would allow the model to learn more effectively from limited data and better adapt to different conditions and subjects. Integrating attention mechanisms into the network would allow computational resources to focus on the most relevant regions of the face during micro-expression detection. This could improve the

accuracy and robustness of detection by concentrating on the most significant details of facial expressions.

Developing methods to detect and classify micro-expressions in more complex environments, such as cluttered scenes or difficult lighting conditions, is essential. This could include algorithms robust to light variations and able to distinguish facial expressions in various contexts.

Studying and implementing data preprocessing techniques to reduce noise and improve the quality of captured events is crucial. Better data preprocessing could lead to more accurate micro-expression detection by eliminating interference and enhancing important features.

8 Conclusion

This project has demonstrated the significant advantages of using event-based vision systems for detecting and analyzing micro-expressions. By integrating event cameras with advanced machine learning algorithms, we have achieved notable improvements in both the accuracy and processing speed of micro-expression recognition compared to traditional methods. Our approach leverages the high temporal resolution and efficiency of event cameras, making it well-suited for real-time applications and environments with varying lighting conditions.

Throughout this research, we have underscored the potential of spiking neural networks (SNNs) in processing the dynamic and temporally sensitive data provided by event cameras. The combination of these technologies offers a robust framework for capturing subtle facial movements that are critical for interpreting human emotions.

In conclusion, this project lays the groundwork for future advancements in micro-expression recognition, offering promising directions for enhancing human-computer interaction, security, and psychological research. Our findings highlight the importance of innovative approaches in overcoming the limitations of traditional methods, paving the way for more nuanced and accurate emotion detection technologies.

References

[1] Berniconi, L., Cultrera, L., Albisani, C., Cresti, L., Leonardo, A., Picchioni, S., Becattini, F., & Del Bimbo, A. (n.d.). Neuromorphic Event-based Facial Expression Recognition. *arXiv preprint arXiv:2304.06351*. Retrieved from https://doi.org/10.48550/arxiv.2304.06351

[2] Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., & Scaramuzza, D. (2022). Event-Based Vision: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1), 154–180. https://doi.org/10.1109/tpami.2020.3008413

[3] Barchid, S., Allaert, B., Aissaoui, A., Mennensson, J., & Djebabra, C. (2023). Spiking-FER: spiking neural network for facial expression recognition with event cameras. *arXiv preprint arXiv:2304.10211*. Retrieved from https://doi.org/10.48550/arxiv.2304.10211

[4] NEFER Dataset. Retrieved from https://www.kaggle.com/datasets/msambare/fer2013/data.

[5] FES (Faces in Event Stream) Dataset. Retrieved from https://www.kaggle.com/datasets/msambare/fer2013/data.

[6] CASME II (Chinese Academy of Sciences Micro-expression) Dataset. Retrieved from https://paperswithcode.com/dataset/casme-ii.

[7] Nexdata/57 Types of Micro-expression Data. Retrieved from https://huggingface.co/datasets/Nexdata/57_Types_of_Micro-expression_Data.

[8] A survey of micro-expression recognitionhttps://www.sciencedirect.com/science/article/abs/pii/S0262885620301750X#:~:text=Micro-expression%20is%20a%20very%20brief%20and%20involuntary%20form,be%20recognized%20because%20of%20their%20subtleness%20and%20brevity..