Understanding Human Reactions Looking at Facial Microexpressions With an Event Camera

Federico Becattini, Federico Palai, and Alberto Del Bimbo, *Senior Member, IEEE*

***Abstract*—With the establishment of *Industry 4.0*, machines are now required to interact with workers. By observing biometrics they can assess if humans are authorized, or mentally and physically fit to work. Understanding body language, makes human–machine interaction more natural, secure, and effective. Nonetheless, traditional cameras have limitations; low frame rate and dynamic range hinder a comprehensive human understanding. This poses a challenge, since faces undergo frequent instantaneous microexpressions. In addition, this is privacy-sensitive information that must be protected. We propose to model expressions with event cameras, bio-inspired vision sensors that have found application within the Industry 4.0 scope. They capture motion at millisecond rates and work under challenging conditions like low illumination and highly dynamic scenes. Such cameras are also privacy-preserving, making them extremely interesting for industry. We show that using event cameras, we can understand human reactions by only observing facial expressions. Comparison with red-green-blue (RGB)-based modeling demonstrates improved effectiveness and robustness.**

***Index Terms*—Biometrics, emotion recognition, event camera, human–machine interfaces, Industry 4.0, microexpressions, neuromorphic sensor, privacy-preserving.**

I. Introduction

ARTIFICIAL intelligence and computer vision have been progressing at a fast pace in the last decade. Such technologies are becoming more applicable in everyday life and in work environments. This development is closely entwined with the birth and growth of what is known as Industry 4.0; a paradigm that puts the interaction between human workers and machines at the center of industrial processes. On the one hand, humans are being equipped with wearable devices and digital identities that facilitate the production process; on the other hand, work environments are becoming smart and interactive. Automation plays an important role in the development of modern workplaces, ranging from robotics to automatic recognition software. Such innovations are raising production and income for industries; yet, they do not come without drawbacks.

A massive amount of confidential or sensible data are being generated. Security of both data and restricted workplaces are necessarily being entrusted to biometric-based algorithms. Workers are now dealing with intelligent machines that could be required to actively recognize or interact with the user, preventing interaction with unauthorized personnel. Even if the machine is able to effectively recognize the worker, some applications might require it to work side-by-side with humans, understanding basic commands, or interpreting feedback, either spoken or provided visually. Therefore, in Industry 4.0, a machine must be able to recognize coworkers, actively interact with them, and understand them by listening or observing. Computer vision is playing a large role in such interaction, since it allows to provide basic forms of intelligence to the machine through sensors such as red-green-blue (RGB) cameras. Whether it is to guarantee safety, restrict access to unwanted users, or simply query a coworker to gather some feedback, relying on vision sensors is fundamental to allow a robotic or digital worker to fulfill its duty. Nonetheless, computer vision still has some limits. In particular, in certain conditions such as scarce illumination or in presence of fast movements, vision methods that are known to perform well in controlled environments fail to provide useful outcomes. This is particularly true when it comes to observing and analyzing faces, which are arguably the most predominant biometric indicators. Faces can undergo huge changes and are highly deformable. What is even more important, especially for security-sensitive applications, is that moods and intentions can be inferred through facial expressions. Nonetheless, the appearance of such underlying emotions can be concealed to an observer, and this can pose issues, if not even threats, to automatic security algorithms.

However, a large crop of literature supports the fact that humans express their true intentions and feelings via subtle and extremely fast movements, known as microexpressions. Such movements of face muscles are involuntary and happen within a few milliseconds. It has been quantified that a microexpression can manifest itself in its entirety between 1/25 and 1/5 of a second [10]. RGB cameras, which normally operate at maximum 25/30 frames per second (FPS), will intrinsically fail at capturing such expressions completely. They might not even be able to capture any signal at all for very fast movements.

However, new types of sensors have proven capable of reducing or even solving some of the aforementioned limitations. Neuromorphic sensors [13], also known as event cameras, have

Manuscript received 15 March 2022; revised 2 June 2022 and 11 July 2022; accepted 26 July 2022. Date of publication 29 July 2022; date of current version 30 September 2022. This work was supported by the Italian MIUR within PRIN 2017 under Grant 20172BH297: I-MALL—improving the customer experience in stores by intelligent computer vision. Paper no. TII-22-1087. *(Corresponding author: Federico Becattini.)*

The authors are with the Media Integration and Communication Center (MICC), University of Florence, 50121 Florence, Italy (e-mail: federico.becattini@unifi.it; federico.palai1@stud.unifi.it; alberto.delbimbo@unifi.it).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2022.3195063.

Digital Object Identifier 10.1109/TII.2022.3195063

been shown to drastically reduce motion blur and have a much higher dynamic range than standard RGB cameras. This kind of sensor has already been used in industrial applications, e.g., high speed counting.[1] A peculiarity of event cameras is the ability to capture motion at rates that dramatically exceed regular cameras, even at a microsecond granularity. This proves useful when highly variable scenes are observed. Despite this, there has not been any attempt in literature to capture the fast motion of a face relying on neuromorphic sensors. In this article, we propose to use event cameras to study facial microexpressions. In particular, we want to automatically understand how humans perceive observed items and how their faces react to reflect such feelings.

An additional challenge for the realization of fruitful Industry 4.0 environments, is to be able to monitor the safety of human workers thanks to their artificial counterparts. Biometric data is continuously collected to understand stress indicators and correct workloads, and avoid any form of harm due to the incorrect use of machinery. Physical and emotional states of the employees are logged for their own safety. Our work also follows this direction, since these kinds of indicators can be derived from a careful analysis of facial expressions. However, privacy concerns are to be expected. The collected data is sensitive, especially when it is vision based, such as photos or videos, and must not be exposed to unregulated access. Ahmad et al. [1] have shown that event-camera data are more privacy-preserving than RGB when collecting biometrics for reidentification. This further motivates our choice of relying on neuromorphic sensors to look at faces, since less privacy issues can rise.

To summarize, the main contributions of this article are as follows.

1) We propose to analyze facial reactions using event-camera data, driven by the intuition that facial microexpressions manifest themselves at rates that make them almost invisible to standard RGB cameras. We show that by relying on the signal produced by an event camera at high speed rates, we are able to better understand the underlying sentiment of observed faces. As a plus, event-camera data are privacy-preserving, further motivating the usage of such kind of sensor.
2) We collect the event-reaction dataset, a set of synchronized RGB and event-camera videos of faces representing human reactions. Each video is paired with a manually labeled reaction score (negative, neutral, positive), provided by the observed user itself.
3) We exploit an event camera simulator to transfer bounding box annotations onto event camera footage. This allows us to train a face detector model without the need of a costly data annotation campaign, bridging the gap between well known computer vision techniques applicable for RGB data and event-based applications.

The rest of this article is organized as follows. In Section II, we frame our work in the current state-of-the-art, discussing relevant topics for Industry 4.0, biometrics, and applications of neuromorphic sensors. To favor a better understanding of the contents of the article, we also provide a brief explanation of how an event camera works in Section III. In Section IV, we provide a definition of the problem we address in this article, i.e., understanding human reactions by looking at microexpressions with an event camera. Then, we explain our data collection process, along with the statistics of the proposed dataset in Section V. Then, we outline our method in Section VI, detailing our face detection model (Section VI-A), the data representation techniques that we exploit to work with event streams (Section VI-B), our classification model (Section VI-C), and its training details (Section VI-D). Experimental validation is reported in Section VII, followed by ablation studies in Section VIII to establish the importance of the components in the model. Finally, Section IX concludes the article.

II. Related Work

A lot of work has been done to equip industrial machinery with an intelligent component. This allows to automate several processes along the manufacturing line and within work environments. In fact, fields such as robotics and artificial intelligence can provide an essential aid into optimizing production pipelines or even taking care of workers with active assistance or assessing their physical and mental state [29].

The role of biometrics is essential. In particular, behavioral biometrics, such as gestures, can provide estimates about the attention status of workers [24] and prevent accidents in the workspace or can help monitoring mental fatigue, which has a direct link with productivity [35]. Among all biometrics, faces play a central role for advanced human–machine interfaces, since they can easily allow reidentification [37] and can express emotional states through expressions [20]. Nonetheless, such microexpressions are fast and difficult to detect [10] and pose the problem of collecting and storing highly sensitive information from a privacy-preservation point of view [34].

Affective recognition in industry and robotics is not limited to analyzing faces and has been largely studied in literature [31]. The problem of recognizing human emotions has often been declined as a module for human–robot interaction applications in order to enhance the sociality of robots [33]. However, this has entailed ethical questions regarding such interactions [9], analyzing the effect that working alongside robots can have on humans [2].

In order to make human–robot interaction effective, several data acquisition modalities might be involved. Liu et al. [19] rely on facial analysis, also exploiting depth and audio signals, while Hong et al. [15] exploit body language captured from a Kinect sensor plus vocal intonation to infer emotional states. Speech alone can also be used [26]. Less common types of sensors have also been tested, such as thermal cameras [12].

Recently, neuromorphic cameras have opened up to new possibilities for computer-vision applications [16], [17], [21], [22]. Such event-based sensors eliminate several limitations of RGB cameras, achieving surprisingly high temporal resolutions and not suffering of motion blur. Furthermore, it has been shown that event cameras are capable of offering a certain

[1][Online]. Available: https://www.prophesee.ai/2019/09/19/high-speed-counting-event-based-vision/

degree of privacy-preservation when employed in biometrics applications [1], since only illumination changes are detected, capturing motion rather than appearance.

Event cameras have already found some use in biometrics and industry, also offering new ways of reidentification, for instance, exploiting eye blinks [7]. Face-centric applications have also been developed, in particular focusing on face detection [25]. Unfortunately, a drawback of using events rather than pixels is the lack of annotated datasets that are extremely costly to produce. This makes it harder to deploy computer vision modules that are of common use with RGB data. To bridge this gap, an event-camera simulator has been proposed, which generates synthetic streams of events from regular videos [27]. In this work, we exploit event camera simulator (ESIM) to train a face detector in the event domain as a first step of our face analysis pipeline.

Despite all the desirable properties of neuromorphic sensors for industrial applications, no prior attempt to exploit its high-speed rates has been made to understand facial expressions. In this article, we make a first step in this direction by learning to recognize human reactions towards observed items. We believe that this will open up to interesting industrial applications and a better comprehension of the human body language. The inspiration of this work derives from a recent study [4], where a similar experiment is performed using only RGB frames. The authors train a classifier to predict the degree of interest towards fashion items in order to perform interactive fashion recommendation. In our article, we extend this idea to the event-based domain by collecting a novel dataset with synchronized RGB-event streams and showing how by using event-camera data we can obtain more reliable and effective predictions.

III. Event Camera

Event cameras are bio-inspired vision sensors, also known as neuromorphic sensors. Such devices, unlike traditional RGB cameras, do not emit a synchronous steam of frames but rather a flow of spatio-temporally localized events. An event is a local change in illumination, that can happen at the extremely high temporal resolution around the order of microseconds, with very low latency. Illumination changes are based on log-illumination, and thus, exhibit a high dynamic range—140 dB, compared to the 60 dB or regular cameras. Each detected event is represented with a polarity depending on the sign of the illumination change. An advantage of neuromorphic sensors is that if no change is detected, no information is produced, thus, reducing bandwidth. Overall, events are generated asynchronously when the illumination change is perceived by the camera. Each event is a tuple (x_i, y_i, t_i, p_i), where x_i and y_i are the spatial coordinates of the corresponding pixel, t_i the event timestamp, and $p_i \in \{-1, 1\}$ its polarity. In order to be processed, events must be either dealt with specialized architectures such as spiking neural networks [14] or converted into frame-based representations [16], [17], [21], [22]. These representations aggregate events happening in accumulation intervals, typically in the order of a few milliseconds.

IV. Problem Statement

In this article, we aim at recognizing human emotions by observing facial expressions with an event camera. In particular, we are interested in quantifying the valence of a reaction to a stimulus perceived by the subject. This can find usage in a wide variety of applications, ranging from understanding a feedback or emotional state of a human worker in an industrial manufacturing process to modeling customer satisfaction for in-shop or online retail. We believe that such applications can also serve as a preliminary study towards understanding true human emotions, beyond any attempt to conceal them.

To establish an experimental setting, we propose to classify human reactions to observed items. Following [4], we use fashion items, for which a vast collection of data is available online. Thus, we formalize the task of understanding human reactions in the following way.

Let $o_i = (t, b)$ be a pair of top and bottom garments, paired to compose an outfit. We define as $s_i^u \in \{-1, 0, 1\}$ the feedback score provided by a user u after observing the outfit o_i. A score of 1 corresponds to a positive feedback, meaning that the user has appreciated the observed outfit. On the contrary, a score of –1 indicates a negative feedback and a score of 0 a neutral feedback. Along with such score, we refer to the video footage of user u observing o_i as v_i^u. The reaction score s_i^u is provided by u right after the footage is collected. Videos can refer to different modalities, either collected with RGB cameras or with event cameras.

Therefore, the task we propose in this article is defined as predicting the reaction score of a user by only observing its reaction footage v, without any additional knowledge of what the user is observing.

Note that the choice of relying on fashion items is purely dictated by data availability and that the prediction model is not tied to the fashion domain in any way, since it only observes facial expressions. Furthermore, we preferred to use such data since it will likely not yield to overempathized reactions. This forces a prediction model to look for subtle microexpressions that are not always voluntary or easily detectable with traditional RGB data.

V. Data Collection

In order to perform our experiments, we collected a novel dataset named event-reaction dataset. This dataset contains videos of subjects reacting to observed fashion outfits. Each sample is obtained by showing an outfit to a user, which is recorded for 10 s. The recording is performed both with an RGB camera and with an event camera. The two sensors are positioned side-by-side and are temporally synchronized. After the 10 s, the user is asked to provide a personal evaluation of the outfit. Such evaluation is provided through a slider, showing three indicators corresponding to a negative, neutral, and positive feedback.

In order to obtain a more fine-grained annotation, we record the slider position as a value in $[0, 100]$ and successively quantize it into the three categories using the values 40 and 60 as thresholds. We leave the positive and negative bins larger in order to include the votes of more conservative users, who tend to

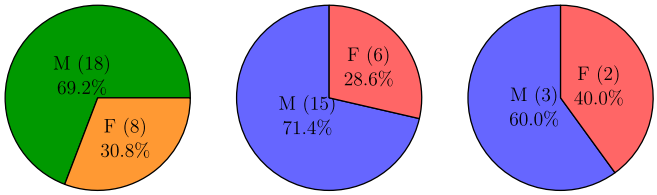

Fig. 1. Dataset gender statistics. On the left, we show the percentage of male and female users in the whole dataset. In the center and on the right, the statistics relative to train and test set.

TABLE I
REACTION STATISTICS IN THE EVENT-REACTION DATASET

	Train	Test	Total	Male	Female
Negative	169	33	202	128 (42.66%)	64 (44.14%)
Neutral	74	9	83	55 (18.33%)	28 (19.31%)
Positive	127	143	170	117 (39.00%)	53 (36.55%)

We collected a total of 455 videos in both the RGB and event modalities, divided into negative (202), neutral (83), and positive (170) reactions, according to user evaluations.

not provide extremely polarized feedbacks and are thus biased towards the neutral class. However, users are unaware of the actual numeric scores that are collected from the slider and use as only reference the three category labels indicating the degree of appreciation.

We collected from 25 different users a total of 455 reactions, which we divided as 80% for training and 20% for testing. Each user has recorded between 15 and 20 reactions. Train and test division is done in order to avoid the same user to be present in both splits. The RGB videos are collected with a camera at a resolution of 640×480 and the Event Camera footage is obtained with a Prophesee GEN 3S video graphics array (VGA)-CD, which has the same resolution of the RGB camera.

The outfits that are shown to the users are drawn at random from a set of prebuilt outfits, taken from the IQON 3000 dataset [32]. Users have an age ranging between 23 and 77 and are distributed as shown in Fig. 1. Statistics on the collected reactions, instead, are provided in Table I. Interestingly, male and female reactions have similar reaction distributions (positive: $\approx 40\%$, neutral: $\approx 20\%$, negative: $\approx 40\%$), meaning that no evident gender bias is present in the data.

VI. METHOD

To understand reaction scores based on facial expressions, we first perform face detection to remove unwanted background noise. We then feed the sequence of cropped faces to a convolutional neural network (CNN) followed by a long short term memory (LSTM), which accounts for the temporal dimension. We perform this operation for either the RGB or event streams. Whereas for RGB this is trivial relying on standard computer vision architectures, for event-based data an event representation has to be chosen.

In addition, we rely on an event-data simulator to train a detector in order to detect faces in event streams. In the following, we outline in detail our data preprocessing pipeline and present our approach to classify user reaction scores, also providing some simpler baselines to compare our model with.

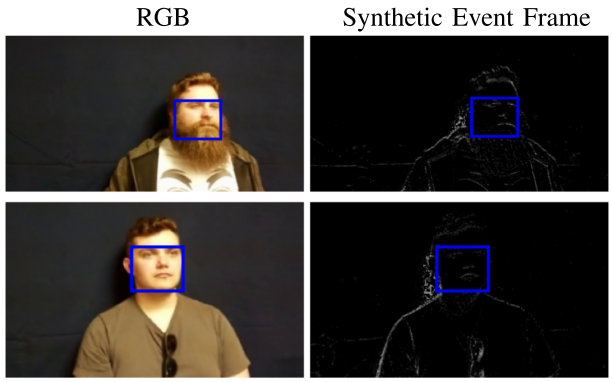

Fig. 2. Using ESIM [27], an event-camera simulator, we generated synthetic event data on which we transferred bounding boxes of faces. Boxes are first obtained on RGB data using face alignment [5].

A. Face Detection

In order to extract faces from RGB videos, we run face alignment [5] on every frame. However, for event-based streams there are no open-source face detectors available, despite some prior work exists in literature [18], [25]. To overcome this limitation, we train our own face detector, relying on ESIM [27], an event-camera simulator.

ESIM generates a synthetic event-based counterpart of an RGB video by using an adaptive sample rate based on the predicted dynamics of the visual signal. Images are rendered at high frame rate, interpolating pixel brightness along the camera trajectory. Reconstructed event frames are generated according to an exponential time surface [17].

We feed to the simulator all the RGB frames to generate a synthetic event-based version of each stream. In this way, we are able to associate the bounding boxes provided by face alignment on RGB frames with event data. We extend the data we collected with data from the Tufts face database [23], a dataset comprising 112 20-s videos of subjects of different ages and ethnicities. In each video a single person is recorded from multiple angles. For these videos we apply the same procedure for generating synthetic bounding-box annotated event data. In Fig. 2, a few samples of the obtained data are depicted, showing the bounding box detected on the RGB frame transferred on the synthetic event-data representation.

Thanks to the combination of face alignment and ESIM, we were able to gather approximately 2 000 000 synthetic frames, using an event accumulation time of 0.001 s. To avoid redundancies we subsampled the data, taking one frame out of ten and, then, we trained a YOLOv3 object detector [28]. On a 20% held-out validation set, our face detector achieved a 96% average precision.

In principle, any frame-based object detector could be trained for our purpose. However, to the best of our knowledge, there is no evidence in literature that detectors based on handcrafted features, such as HOG [3], [11] or LBP [36], could be adaptable to event data. Therefore, we preferred to rely on a learning-based state-of-the-art detector such as [28].

We use such face detectors to locate faces in real event camera streams for our event-reaction dataset. This is possible since the

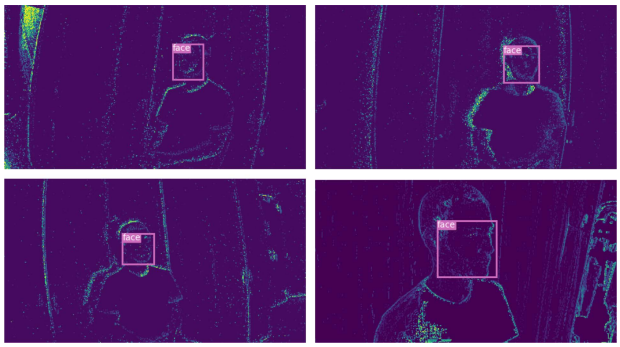

Fig. 3. Detections on real data acquired with an event camera. The detector is trained on synthetic data and applied on real events.

synthetic streams on which the detector is trained are sufficiently similar to real event-based data. A *synth-to-real* domain shift has been studied [6], [27], identifying as main difficulties the difference in recording settings and the simplistic noise model adopted by the simulator. However, for the purpose of training a face detector, we did not observe any relevant complication when transitioning from the simulated domain to the real one. Examples of detections on real event frames are shown in Fig. 3.

Since all videos contain a single user, we first link all boxes across adjacent frames using a box association tracker [8] and, then, take the object with the highest average box score. For missing detections, e.g., when no movement is recorded, we interpolate the coordinates of the box in the previous and following frames. This approach has two advantages: 1) it is easily adaptable to videos with multiple users, since [8] is a multitarget tracking-by-detection algorithm; 2) it operates on mid-level representations (i.e., bounding boxes), thus, it does not require specific training for different data sources, making it suitable for both RGB and event data without additional annotations. Event-based trackers [30] could have been used directly on raw data, but, as for the object detector would have required labeled face data and would have been specific for the event domain. Furthermore, we collected the data in a constrained environment with low clutter and no camera motion, which makes box association trackers such as [8] capable of providing near perfect tracking results.

B. Data Representation

To represent event camera data, we use three different aggregation functions from the state-of-the-art: 1) temporal binary Representation (TBR) [16]; 2) surface of active events (SAE) [21]; and 3) polarity [22].

We need an aggregation function since we rely on frame-based architectures taken from standard RGB computer vision approaches to address our task.

Temporal binary representation is an encoding based on two steps. First, temporal binary slices are obtained, by checking for presence or absence of any event in a small time interval Δt. Then, a fixed number of slices N is grouped together by converting for each pixel the sequence of binary digits into a base-10 value. We use a $\Delta t = 10$ ms and $N = 8$. Therefore, the final frame-based representation F_{TBR} is obtained as

$$F_{TBR} = bin2dec(B_0 \ldots B_N) \tag{1}$$

where B_i is the ith binary slice.

SAE, instead, measures the time interval occurring between observed events and a starting point t_0. Each frame refers to an observation interval Δt, which we set to 10 ms. Similarly to TBR, polarity is discarded, whereas values are scaled in [0, 255]. The frame representation F_{SAE} is thus obtained as

$$F_{SAE}(x, y) = 255 \times \left(\frac{\text{tp} - \text{t0}}{\Delta \text{t}} \right) \tag{2}$$

where t_p is the timestamp of the last observed event at a given pixel.

Finally, polarity encodes values according to the polarity of the most recent event, namely using 0 if the event has negative polarity, 1 if it is positive, and 0.5 if no event happens in the accumulation time interval. Therefore, the final representation is a frame F_{POL}, where each pixel (x, y) is given by

$$F_{POL}(x, y) = \begin{cases} 0, & \text{if event polarity is negative} \\ 0.5, & \text{if no event is observed in } \Delta \text{t} \\ 1, & \text{if event polarity is positive.} \end{cases} \tag{3}$$

C. Reaction Classifier

We build our reaction classifier as a frame-based convolutional network. Since we deal with videos, we need to learn a spatial representation to capture important facial features and then model the temporal dimension. We exploit a CNN backbone which is applied in a time-distributed fashion, i.e., applying it for each frame in the video sequence. The backbone is built as a sequence of four convolutional layers followed by ReLU activation and max-pooling. Then, the final feature map is flattened and fed to two fully connected layers, yielding a compact 128-dimensional feature vector. Then, two LSTM layers process these sequences of features, providing the final classification thanks to two readout dense layers and a softmax activation. Fig. 4 depicts our model. In the following we will refer to this model as CNN+LSTM.

We also provide two baseline models. First, we train the backbone model without the temporal part, i.e., truncating the model at the first fully connect layer and adding a three-way classification head. We train it to classify only the middle frame of each video, thus, only relying on spatial information alone. This model serves the purpose of highlighting the importance of modeling the temporal dimension for the task at hand. We refer to this model as CNN-mid.

Then, we propose a baseline that still exploits the same architecture of CNN-mid, but is applied to each frame of the video. The dense features generated for each frame by the first fully connected layer are then max-pooled together before passing them to the classification head. This model offers a simpler way to account for temporal dynamics, without relying on specifically tailored temporal models. We refer to this model that only uses the convolutional backbone followed by a max-pooling as CNN+MP.

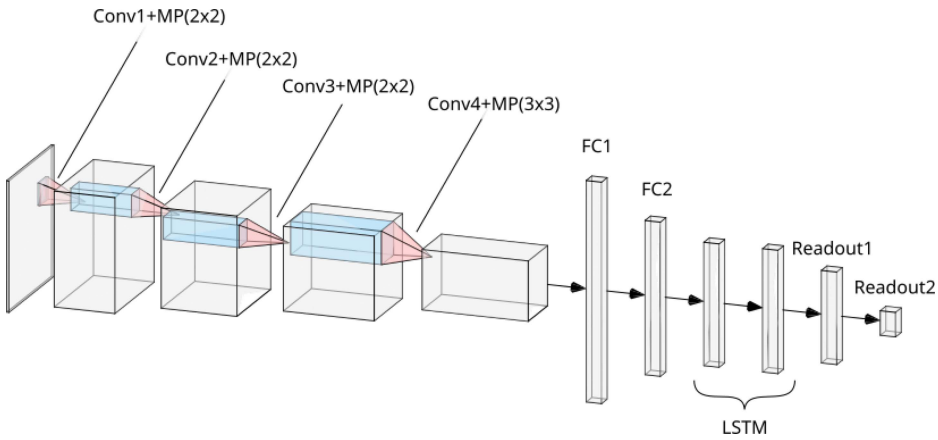

Fig. 4. Model architecture. First, a convolutional backbone processes input frames. The resulting feature maps is flattened and passed through two fully connected layers before being fed to a recurrent model composed of two LSTM layers. The final classification is performed by two readout fully connected layers and softmax activation.

D. Training Details

In this section, we provide training and architecture details to foster reproducibility.

First of all, input images for each model are resized to 267 × 367 pixels. In the CNN+LSTM model, the convolutional layers use all 3 × 3 kernels with an increasing number of channels: 16, 32, 64, 128. The two fully connected map the feature to latent spaces of dimension as 1024 and 128, respectively, and use dropout during training. The two LSTMs instead have a hidden dimension of 256 and the final readout layers project the output first to 128 dimensions and finally to 3, corresponding to the negative, neutral, and positive classes.

All models have been implemented in PyTorch and have been trained for 30 epochs on an Nvidia Titan RTX GPU. We used the Adam optimizer with learning rate of 0.0001 and a cross-entropy loss. Since the collected data exhibits class unbalance, we assign class weights to the loss using the following criterion. We assign to class i the weight $w_i = \frac{C}{c_i}$, where c_i is the number of samples in class i and C is the number of samples for the most populated class. In this way, under-represented classes get a bigger update during the backpropagation step.

VII. Experimental Results

We evaluate our experiments on the collected dataset in Table II. We show accuracy results for the three proposed models CNN-mid, CNN+MP, and CNN+LSTM. We also report a random baseline to provide a lower bound on the results. Since the dataset is unbalanced with reference to the gender of the users, along with the accuracy averaged over all samples, we also show an accuracy breakdown by gender.

Each model is evaluated using four different variants. The first variant uses RGB frames as input, while the others rely on real event camera data, aggregated according to polarity [22], SAE [21], or TBR [16], as outlined in Section VI-B. Note that, here, we do not use synthetic events but only real data collected as explained in Section V.

TABLE II
Accuracy of the Proposed Methods With Different Input Data

Input data	Method	Accuracy		
		Average	Male	Female
-	Random	25.88	25.79	26.01
RGB	CNN-mid	27.16	27.07	27.29
E-Polarity	CNN-mid	27.39	27.38	27.41
E-SAE	CNN-mid	27.81	27.84	27.77
E-TBR	CNN-mid	28.14	28.05	28.27
RGB	CNN+MP	29.71	29.68	29.75
E-Polarity	CNN+MP	31.87	31.90	31.83
E-SAE	CNN+MP	32.66	32.74	32.53
E-TBR	CNN+MP	33.02	33.05	32.98
RGB	CNN+LSTM	43.50	43.52	43.47
E-Polarity	CNN+LSTM	44.93	45.02	44.79
E-SAE	CNN+LSTM	45.26	45.19	45.36
E-TBR	CNN+LSTM	**45.88**	**45.85**	**45.92**

We compare RGB frames against event-based representations: polarity [22], SAE [21], and TBR [16].

First of all, it can be seen that modeling the temporal dimension is of primary importance for successfully addressing the task. All the variants of the CNN-mid model achieve an accuracy slightly above random, thus, failing to produce meaningful results. Considering a single frame is not enough to correctly understand the reaction of the user.

When including time through pooling, i.e., using the CNN+MP model, the recognition rate slightly improves. Nonetheless, a pooling function discards a lot of information and loses the temporal ordering of frames. Interestingly, the event-based representations are able to provide better results even when time is provided out of order via a pooling function.

However, when explicitly modeling the temporal dimension with an LSTM, we report a big improvement in accuracy. All methods register and increase of more than ten points, with the TBR event-based model achieving the best results. It is interesting to observe that the model struggles more with RGB inputs. The three event aggregation strategies report improvements of 3.28% (polarity), 4.04% (SAE), and 5.47% (TBR), indicating

Fig. 5. Samples of temporal binary representation frames obtained for samples in the event-reaction dataset and using $\Delta t = 10$ ms and $N = 8$ bits. Each frame covers an overall timespan of 80 ms. Data is acquired with an actual event camera (no simulator involved). Image is best viewed in color on computer screen.

that event-based representations are a rich and informative way to encode facial expressions.

We impute this to the ability of event cameras to capture movements at an extremely high frequency. In fact, RGB video being recorded at 20 FPS provide snapshots every 50 ms, whereas the event-based representations that we used produce a frame every 10 ms. This comes at a cost in complexity, since more frames have to be processed, but underlines the importance of a fine-grained temporal analysis. In fact, facial expressions are very subtle and can happen at rates between 1/25 s and 1/5s [10]. If regular cameras may capture such facial movements, they do it partially, without observing the phenomena in its entirety.

Additionally, it is worth to note that among the three adopted event aggregation strategies, TBR, which achieves the best results, actually encodes temporal information for N slices together in the same frame. This has two important consequences: 1) each pixel carries both spatial and temporal information; 2) when keeping the accumulation time fixed, the number of frames to be processed is halved by a factor of N. Using $\Delta t = 10$ and $N = 8$, we are in fact generating frames that condense information at a 10-ms granularity yet generating only 12.5 frames per second. This is 1.6 times lower than the number of frames to be processed for RGB and 8 times lower compared to the other event-based representations. On average, on an Nvidia Titan RTX it takes 43 ms to process RGB sequences, 27 ms to process sequences encoded with TBR, and 108 ms for sequences encoded with SAE or polarity. We believe the number of frames is the main cause of the slight decrease in performance of SAE and polarity compared to TBR, since the LSTM layers have to process much longer sequences. Examples of TBR representations for samples in the gesture-reaction dataset are shown in Fig. 5.

TABLE III
ABLATION STUDY

Input data	Face Detection	Accuracy
RGB	✓	43.50
E-TBR	✓	45.88
RGB	✗	41.73
E-TBR	✗	39.11

Importance of using ROIs of detected faces instead of full frames. All methods are trained using the CNN+LSTM model.

Furthermore, event-based data, compared to RGB, helps the model to focus on relevant parts of the image that are active during an expression. In fact, if parts of the face remain still, the corresponding pixels will be empty. On the contrary, when a facial expression is being performed, there will be strong activations in the frame, guiding the network towards relevant regions. In Fig. 6, we show sample crops from an RGB video and its correspondent event-based stream.

VIII. Ablation Studies

We now perform a series of ablation studies to assess the importance of three components: 1) using the object detector compared to the full frame; 2) using different loss weights to avoid class unbalance; 3) the effect of stacking LSTM layers in the model.

First, we report in Table III the results for the CNN+LSTM model with and without using face crops. We show both accuracies for RGB and TBR data. In both cases we can observe a drop in accuracy. However, the drop for TBR is much higher than the one registered for RGB. This can be imputed to excessive background noise that makes the subtle movements of the face

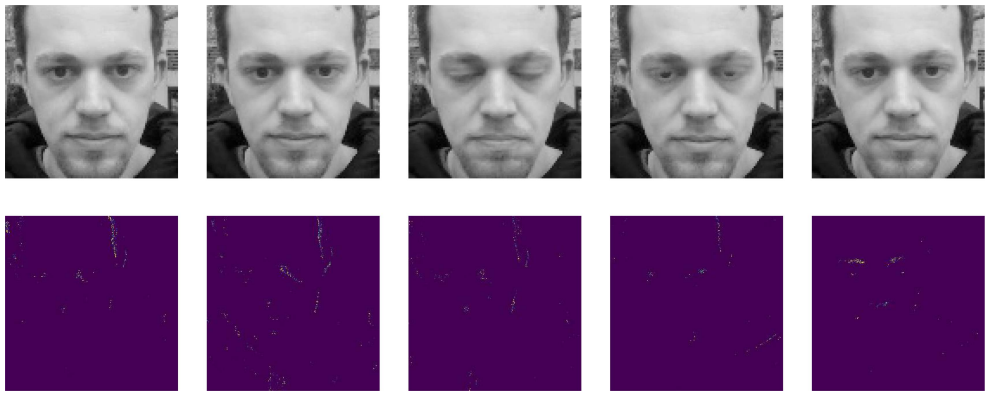

Fig. 6. Cropped faces in RGB and in the event domain. RGB frames capture some of the expressions, without providing a sufficient temporal granularity. Event based data captures only moving parts involved in the expression, such as lips, cheeks, eyes, and eyebrows. Data are acquired with an actual event camera (no simulator involved). Image is best viewed in color on computer screen.

TABLE IV
ABLATION STUDY

Input data	Weight	Accuracy
RGB	$\frac{C}{c_i}$	43.50
E-TBR		45.88
RGB	$\frac{1}{c_i}$	41.84
E-TBR		44.96
RGB	–	35.29
E-TBR		44.70

Note: Importance of weighting classes during training. All methods are trained using the CNN+LSTM model. Class weights are based on class cardinalities c_i, with $C = \max_i\{c_i\}$.

TABLE V
ABLATION STUDY

	Num layers		
Input data	1	2	3
RGB	42.98	43.50	43.16
E-TBR	45.17	45.88	45.39

Accuracy varying the number of LSTM layers in the recurrent part of the CNN+LSTM model.

less evident. Event cameras are in fact quite sensible to environmental noise. At the same time, using the whole frame, we are reducing the surface of the area of interest which includes the face. In RGB frames the signal is downsampled but still visible, whereas for event data the small activations area might get too decimated.

We then study the effect of adding class weights in the cross-entropy loss. As detailed in Section VI-D, we weigh class samples as $\frac{C}{c_i}$, i.e., normalizing the weights to be 1 for the largest class. We also add a second weighing strategy, using as weights $\frac{1}{c_i}$, which corresponds to dividing the loss for the cardinality of the class. In this way, all losses are dampened. In Table IV, we show the results for the three approaches. As expected, without using weights for balancing classes, the accuracy drops considerably, especially for RGB. We also observe that boosting the samples of less-populated classes helps more than penalizing all classes depending on the number of samples.

Finally, we change the number of LSTM layers of the recurrent part of the model. In our standard model, we use two LSTMs with a hidden dimension of 256 each. We try varying the number of layers, using just a single layer or stacking three layers on top of each other. Results are shown in Table V. When using two layers we are able to reach the highest accuracy. We observe a slight decrease when adding a third layer, which is adding too much complexity to the model compared to the limited amount of data used for training. When using just a single layer, instead, the accuracy drop increases, hinting that the model is not able to identify complex patterns as well as before.

IX. CONCLUSION

In this article, we presented a novel application which strived to capture subtle facial microexpressions thanks to the high rate of an event camera. We collected a dataset of video user reactions which do not exhibit overly-emphasized facial movements. On such data, RGB cameras failed to properly record fast micro-movements and we showed that event camera-based streams were instead suitable for addressing these kind of tasks. We demonstrated the capabilities of a spatio-temporal model based on a convolutional backbone followed by LSTM layers. We relied on three different event aggregation strategies to generate frame representations.

We believe that in the near future, with the diffusion of neuromorphic cameras in work environments, applications based on interactions between humans and computer interfaces will become more robust and secure. Fine-grained human understanding requires an analysis that goes beyond the capabilities of standard cameras since certain signals, such as facial microexpressions, can be only partially captured and observed. Analyzing faces and biometrics in general with event cameras paves the way towards interesting and currently unexplored applications, which will find usage along with the Industry 4.0 paradigm as well as more traditional fields like video surveillance.

As future work we intend to extend our analysis by collecting data with new generation event cameras that offer a higher spatial resolution. In fact, a VGA resolution like the one used in this work, poses a limitation towards capturing spatially

fine-grained facial movements, which are of primary importance for facial microexpressions. We also intend to capture a larger dataset involving subjects of different ages, genders, and ethnicities in order to increase the capacity of our models to generalize. In addition, a more fine-grained characterization of reactions would be desirable, labeling emotions instead of positive/neutral/negative reactions.

REFERENCES

[1] S. Ahmad, G. Scarpellini, P. Morerio, and A. Del Bue, "Event-driven Re-Id: A new benchmark and method towards privacy-preserving person re-identification," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 459–468.

[2] N. Akalin, A. Kristoffersson, and A. Loutfi, "Do you feel safe with your robot? Factors influencing perceived safety in human-robot interaction based on subjective and objective measures," *Int. J. Hum.- Comput. Stud.*, vol. 158, 2022, Art. no. 102744.

[3] F. Becattini, L. Seidenari, and A. Del Bimbo, "Indexing quantized ensembles of exemplar-SVMS with rejecting taxonomies," *Multimedia Tools Appl.*, vol. 76, no. 21, pp. 22647–22668, 2017.

[4] F. Becattini et al., "PLM-IPE: A pixel-landmark mutual enhanced framework for implicit preference estimation," in *Proc. ACM Multimedia Asia*, 2021, pp. 1–5.

[5] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1021–1030.

[6] M. Cannici et al., "N-ROD: A neuromorphic dataset for synthetic-to-real domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1342–1347.

[7] G. Chen, F. Wang, X. Yuan, Z. Li, Z. Liang, and A. Knoll, "Neuro-Biometric: An eye blink based biometric authentication system using an event-based neuromorphic vision sensor," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 1, pp. 206–218, Jan. 2021.

[8] G. Cuffaro, F. Becattini, C. Baecchi, L. Seidenari, and A. Del Bimbo, "Segmentation free object discovery in video," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 25–31.

[9] L. Devillers, "Human–robot interactions and affective computing: The ethical implications," in *Robotics, AI, and Humanity*. Cham, Switzerland: Springer, 2021, pp. 205–211.

[10] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*. New York, NY, USA: Norton, 2009.

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[12] C. Filippini, D. Perpetuini, D. Cardone, A. M. Chiarelli, and A. Merla, "Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review," *Appl. Sci.*, vol. 10, no. 8, 2020, Art. no. 2924.

[13] G. Gallego et al., "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.

[14] S. Ghosh-Dastidar and H. Adeli, "Spiking neural networks," *Int. J. Neural Syst.*, vol. 19, no. 4, pp. 295–308, 2009.

[15] A. Hong et al., "A multimodal emotional human-robot interaction architecture for social robots engaged in bidirectional communication," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 5954–5968, Dec. 2021.

[16] S. U. Innocenti, F. Becattini, F. Pernici, and A. Del Bimbo, "Temporal binary representation for event-based action recognition," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 10426–10432.

[17] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman, "HOTS: A hierarchy of event-based time-surfaces for pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1346–1359, Jul. 2017.

[18] G. Lenz, S.-H. Ieng, and R. Benosman, "Event-based face detection and tracking using the dynamics of eye blinks," *Front. Neurosci.*, vol. 14, 2020, Art. no. 587.

[19] Z. Liu et al., "A facial expression emotion recognition based human-robot interaction system," *IEEE/CAA J. Automatica Sinica*, vol. 4, no. 4, pp. 668–676, Sep. 2017.

[20] W. Merghani, A. K. Davison, and M. H. Yap, "A review on facial micro-expressions analysis: Datasets, features and metrics," 2018, *arXiv:1805.02397*.

[21] E. Mueggler, C. Bartolozzi, and D. Scaramuzza, "Fast event-based corner detection," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2017, pp. 33.1–33.11.

[22] A. Nguyen, T.-T. Do, D. G. Caldwell, and N. G. Tsagarakis, "Real-time 6DoF pose relocalization for event cameras with stacked spatial LSTM networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1638–1645.

[23] K. Panetta et al., "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, Mar. 2020.

[24] A. N. Patel et al., "Mental state assessment and validation using personalized physiological biometrics," *Front. Human Neurosci.*, vol. 12, 2018, Art. no. 221.

[25] B. Ramesh and H. Yang, "Boosted kernelized correlation filters for event-based face detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops*, 2020, pp. 155–159.

[26] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, "Speech emotion recognition in emotional feedback for human-robot interaction," *Int. J. Adv. Res. Artif. Intell.*, vol. 4, no. 2, pp. 20–27, 2015.

[27] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. Conf. Robot Learn.*, 2018, pp. 969–982.

[28] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[29] L. D. Riek, "Healthcare robotics," *Commun. ACM*, vol. 60, no. 11, pp. 68–78, 2017.

[30] C. Ryan et al., "Real-time face & eye tracking and blink detection using event cameras," *Neural Netw.*, vol. 141, pp. 87–97, 2021.

[31] J. D. Schwark, "Toward a taxonomy of affective computing," *Int. J. Hum.-Comput. Interaction*, vol. 31, no. 11, pp. 761–768, 2015.

[32] X. Song, X. Han, Y. Li, J. Chen, X.-S. Xu, and L. Nie, "GP-BPR: Personalized compatibility modeling for clothing matching," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 320–328.

[33] M. Szabóová, M. Sarnovskỳ, V.M. Krešňáková, and K. Machová, "Emotion analysis in human–robot interaction," *Electronics*, vol. 9, no. 11, 2020, Art. no. 1761.

[34] Q.N. Tran, B. P. Turnbull, and J. Hu, "Biometrics and privacy-preservation: How do they evolve?" *IEEE Open J. Comput. Soc.*, vol. 2, pp. 179–191, Mar. 2021.

[35] L. J. Trejo, K. Kubitz, R. Rosipal, R. L. Kochavi, L. D. Montgomery, "EEG-based estimation and classification of mental fatigue," *Psychol.*, vol. 6, no. 5, 2015, Art. no. 572.

[36] X. Wang, T. X. Han, and S. Yan, "An HOG-LBP human detector with partial occlusion handling," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 32–39.

[37] D. Wu et al., "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, pp. 354–371, 2019.

Federico Becattini received the Ph.D. degree in information engineering from the University of Florence, Florence, Italy, in 2018.

He is currently a Postdoctoral Researcher and Adjunct Professor with the University of Florence. He has authored or coauthored more than 25 scientific papers in journals and international conferences. His research interests focus on autonomous driving, human behavior understanding, fashion recommendation, and working with international academic and industrial partners.

Dr. Becattini serves as a Reviewer for top-tier conferences and journals in multimedia and computer vision. Recently, he has coorganized the workshops "Facial and Body Expressions (FBE)" at the 25th International Conference on Pattern Recognition (ICPR2020), "Towards a Complete Analysis of People: From Face and Body to Clothes (T-CAP)" at the 21st International Conference on Image Analysis and Processing (ICIAP2021) and ICPR2022, "Multimedia Computing towards Fashion Recommendation (MCFR)" at the 30th ACM International Conference on Multimedia (ACMMM 2022), and "1st International Workshop and Challenge on People Analysis: From Face, Body and Fashion to 3-D Virtual Avatars (WCPA)" at the European Conference on Computer Vision 2022 (ECCV 2022). He has held tutorials on Memory Networks at ICIAP 2022 and ACMMM 2022.

Federico Palai received the master's degree cum laude in computer engineering from the University of Florence, Florence, Italy, in 2021 with the thesis "Event-based Facial Expressions Analysis," under the supervision of Prof. Alberto Del Bimbo and Prof. Federico Becattini.

Alberto Del Bimbo (Senior Member, IEEE) received the master's degree cum laude in electrical engineering, in 1977.

He is a Full Professor of computer engineering, and the Director of the Media Integration and Communication Center, University of Florence, Florence, Italy. From 1996 to 2000, he was the President of the IAPR Italian Chapter and from 1998 to 2000, the Member-at-Large with the IEEE Publication Board. His research interests include multimedia information retrieval, pattern recognition, and computer vision.

Prof. Del Bimbo received the SIGMM Technical Achievement Award for Outstanding Technical Contributions to Multimedia Computing, Communications and Applications. He was nominated the ACM Distinguished Scientist in 2016. He was the General Co-Chair of the 18th ACM International Conference on Multimedia (ACMMM2010) and the 12th European Conference on Computer Vision (ECCV2012). He is an IAPR Fellow, and the Associate Editor for *Multimedia Tools and Applications*, *Pattern Analysis and Applications*, *Journal of Visual Languages and Computing*, *International Journal of Image and Video Processing*, *Pattern Recognition*, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.