Final report

File no: 36-3400.7/100 **Project:** EmbeddedNeuroVision

Project partners: FZI, HSA, Inferics

Number of companies involved: 2 (HSA, Inferics)

Number of SMEs involved included: 2 (HSA, Inferics)

FZI Research Centre for Information Technology (FZI)

Haid-und-Neu-Str. 10-14
76131 Karlsruhe

Victor Pazmino
pazmino@fzi.de
+49 (0)721 9654 190

HS Analysis GmbH (HSA)

Haid-und-Neu-Strasse 7
76131 Karlsruhe

Sergey Biniaminov
sergey.biniaminov@hs-analysis.com
+49 (0)7247 93429980

Inferics GmbH (Inferics)

Benzstr. 3
76185 Karlsruhe

Dr Norbert Link
norbert.link@inferics.com
+49(0)721 721 3528 732

The report is published on the project website:

https://www.embeddedneurovision.de/

and is also made available by the partners at the following Internet addresses:

https://cloud.inferics.com/index.php/s/aWQMqtJg2Z9dHXK

https://www.hs-analysis.com/embnv/

Table of contents

1. Project results

Project goal

The goal of the EmbeddedNeuroVision project was to compare neuromorphic sensors and AI systems with currently used, non-neuromorphic variants. Specifically, in the context of human action recognition, a system consisting of a neuromorphic camera (also called an event-based camera) and neuromorphic AI processing (spiking neural network on neuromorphic hardware) was to be compared with a classic system consisting of an image-based camera and artificial neural network. In addition, it should be investigated whether individual processing steps of the two systems can be exchanged, e.g. whether a classic camera can be used in combination with neuromorphic AI processing by cleverly converting the data.

Milestones

The project work was divided into six work packages whose progress and completion was checked via the milestones listed in Table 1. All work was successfully completed. A brief description of the work and results as well as any deviations from the project plan follow for each milestone after the table.

Table 1: Milestones including processing status.

MS	Description	Status
M0.1	Project website and logo was set up	completed
M1.1	Requirements, interfaces and system architecture defines	completed
M2.1	Integration of conventional camera completed	completed
M2.2	Conversion algorithm implemented	completed
M2.3	Event-based camera integrated	completed
M3.1	Artificial neural networks selected	completed
M3.2	Trained KNN available	completed

M3.3	KNNs were newly developed based on event-based data. trains	completed
M4.1	Spiking Neural Network selected	completed
M4.2	Trained SNN available	completed
M4.3	Evaluation of the SNN completed	completed
M5.1	Evaluation completed, project results available	completed

M0.1 Project website and logo: At the beginning of the project, a suitable logo was designed that combines the main aspects of "vision" and "AI" and thus makes the topics of the project quickly comprehensible. In addition, a website was set up to present the project goals, motivation and results to the public:

https://www.embeddedneurovision.de/

M1.1 Requirements, interfaces and system architecture: In a joint brainstorming session, the system architecture is designed according to the design thinking approach: The basic elements are the event-based camera for

"image" capture and subsequent processing on a computer. Both parts are captured as individual modules and are intended to act autonomously. In contrast to commonly known frame-based cameras, which display temporally discrete, approximately synchronous images of the entire sensor (frames), event-based cameras perceive changes in the charge current of individual pixels of the sensor and thus output a sequence of asynchronous events as a signal. The resulting data stream thus contains information about the location and magnitude of a change, independent of the state of the remaining pixels. This results in other requirements for the design of the overall system: with regard to the real-time of the system, the sampling rate remains relevant, which is measured here in mega-events per second (MEPS) and is thus several orders of magnitude higher than with frame cameras. At the same time, it is still limited in time, so that a decision must be made between frequency and temporal coding (rate coding vs. temporal coding). With rate coding, the number of pulses within a sampling period is recorded, so that a high number of pulses (spikes) provides a stronger signal. Temporal coding records a single spike within the sampling period, whose time of occurrence relative to the sampling period determines the signal strength.

strength of the signal is determined. Rate coding enables the conversion of conventional ANNs, while temporal coding allows smaller, more efficient networks. With regard to the formulated goal of designing an energy-saving product, it is advantageous to use temporal encoding in this project. On the hardware level, on the other hand, we can work with similar requirements as for frame cameras. Here, a standardised USB interface is used as a requirement; for later, more in-depth integration, a MIPI-CSI2 is planned for PCB mounting, so that a normal USB type A interface can be used as an interface to the SNN-processing hardware.

M2.1 Integration of conventional RGB-d camera: The hardware basis is the product PatronuSens Professional from Inferics GmbH, which uses an omnidirectional stereo camera to record both RGB image data and calculate 3D data of the observed half-space, so that it delivers streams of RGB 3D data (for each pixel also the x,y,z coordinates) as well as a stream of 3D coordinates of the keypoints (joints, eyes, ears, nose) of all persons in the room by means of the built-in AI. In accordance with the procedure for processing data from event cameras, in which frames with pixel activations corresponding to the events supplied are created from the event data stream for processing, so-called event frames are created from successive frames of the conventional camera, in which the activation of suprathreshold changing pixels is represented, whereby these can be RGB event frames, 3D event frames or keypoint event frames. The "event frame camera" created in this way thus uses the image streams of its integrated conventional PatronuSens frame camera to generate an event frame data stream from it that qualitatively corresponds to the data as processed from an event camera.

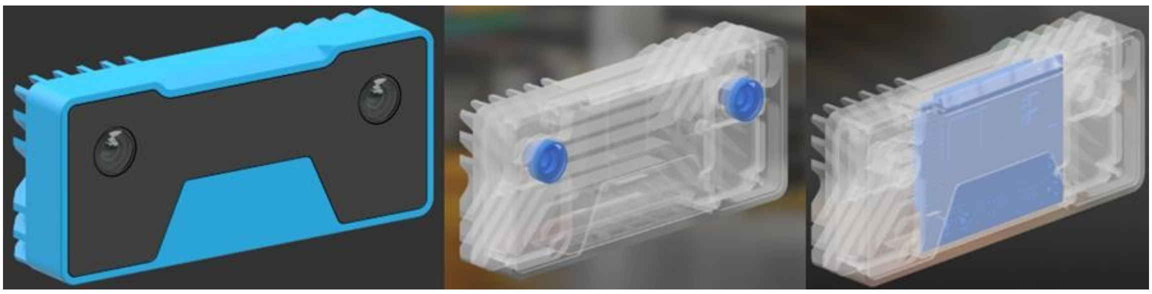


Figure 1: EmbeddedNeuroVision Eventframe Camera Hardware

The event frame camera is realised as a software module that generates the current event frame data from the data of the last two frames of the built-in camera and makes it available to the evaluation module in event frame format via a software interface, as shown in the following diagram.

Figure 2: Architecture of the event frame camera

M2.2 Conversion algorithm for frame camera data streams to event camera data streams and event frame data streams: If there is only one frame camera available, the data streams are converted to event camera data streams.

The temporal correlation can be displayed in accordance with the event frame representation. to approximate the data streams of the event camera. This can be approximated by a frame camera when the colour value changes between two frames are evaluated. Such an emulation of event cameras has already been proposed in the literature, where for each pixel of two consecutive frames a lin-log function (approximation of the logarithm) of the intensity I (formed from the RGB values I=R+G+B) or the colour value h = b/(r+g) is calculated. The difference of these values between two frames is examined for a change with a suprathreshold amount and in the positive case the pixel is assigned the value ON or OFF (depending on the sign of the difference). Since several pixels are assigned ON/OFF values at the same time and many evaluation methods cannot process identical time stamps, the time interval between two frames is linearly interpolated according to the number of simultaneous pixels, and the interpolation values are calculated as as individual timestamps assigned.
Although this established type of emulation leads to data streams that correspond to those of event cameras, they are not processed by the common evaluation methods using convolutional layers and spiking neural networks and do not produce satisfactory results when processed directly with recurrent networks, as was shown in the work. Instead, an aggregation of the events takes place as pre-processing, in that the event coordinates for the

The data streams can be used to create an "event image" or "event frame" in which events of the corresponding pixel coordinate lead to an entry at the corresponding position in the image with the associated polarity. Therefore, a conversion algorithm for data streams from a frame camera was developed in the project, which delivers event frame data streams corresponding exactly to this representation, which can be evaluated analogously to those of the event camera.

The event frame camera thus produces a set Y of pixels of simultaneous events ("event frame") that represent the changes in colour value over a threshold between two image frames: Calculate the difference image of successive frames $I(t)$ and $I(t + \Delta)$ with adjustable frame rate. Determine the amount Y of overshadowed pixels:

$$[u, v, \Delta r, \Delta g, \Delta^{b]T} \in Y, wenn \sqrt{([r, g, b]|_{u,v,t} T - [r, g, b]|_{u,v,t+\Delta} T)^2} > c$$

In this case, enter the values $\Delta r, \Delta g, \Delta b$ at the corresponding pixel position $u, v,$ otherwise the value 0.

M2.3 event-based camera is integrated: The hardware used is the CeleX5_MP development kits from CelePixel Technology Co. LTD and Gen 4.0 from Prophesee. It should be noted that the company CelePixel was bought by Omnivision Sensor Solution Co. LTD in the course of the project, so that this particular kit is no longer available. The publicly available documentation of the control and decoding of the data, which can be found here, serves as the basis for the integration:

https://github.com/CelePixel/CeleX5-MIPI.

The CelePixel camera offers several modes for data acquisition, the one corresponding to Temporal Encoding (described in M1.1) is the Event-In-Timestamp mode. Here, each data point has the structure:

Line [py], column [px], entry time [µs], transmission time [µs].

Thus, the temporal encoding can be taken from a stored file as well as from the continuous data stream by considering the coordinates and the time of occurrence of the events.

In contrast, the data from the Prophesee camera is output in the following format by default, which also contains the polarity of the events (i.e. whether a pixel has become lighter or darker):

Column [px], row [py], polarity [binary], transmission time [µs].

However, polarity was not used in the project for better comparability. For a successful SNN analysis with the present SNN frameworks, the full-picuture mode was used in practice, since a complete matrix is required for the analysis in the subsequent SNN, especially the connection to the neuromorphic hardware. Here, all events in a defined time period are summarised in a matrix with the dimensions of the resolution of the camera and passed on collectively as a frame to the analysis.

M3.1 Development of Artificial Recurrent Neural Networks for Activity Recognition from Frame Camera and Event Frame Data Streams: As a

Baseline was the person activity recognition system previously developed by Inferics.
from data streams of frame cameras, where first for each frame (consisting of a stereo image pair) the 3D coordinates of the 17 body key points (joints and the head key points eyes, ears and nose) are extracted by means of a deep convolutional network, from these a 17x3 pose frame is calculated and sequences of these pose frames are processed with a time convolutional network (t-CNN). This basic method ("Pose+t-CNN") should be contrasted with a new approach based on recurrent networks: For uniform processing of processes of different lengths in time, an "unconventional" echo state network, ESN, (strongly related to spiking neural networks, as used by the partner FZI), which directly represents different time courses by internal states (echo states), should represent an embedding of the time signals in the form of the state expressions, which should then be processed as a sequence of states by a recurrent network of the type LSTM for recognition, as shown in the diagram below.

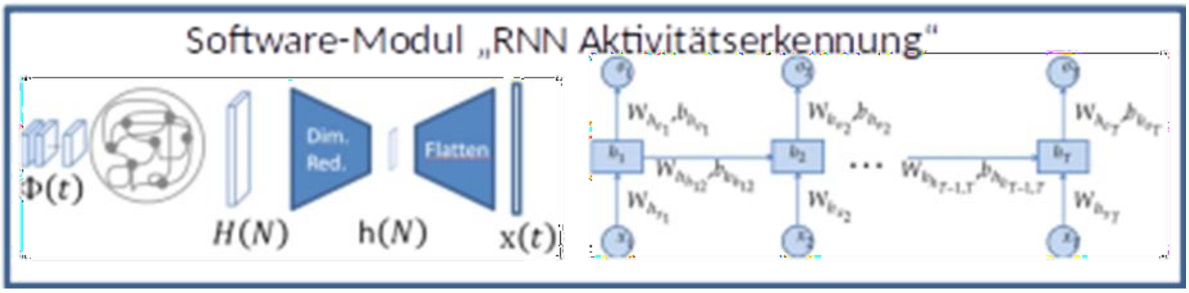


Figure 3: ***Original approach with ESN (left) and LSTM (right)***

However, preliminary studies of the ESN with the simplified case of detecting pathological conditions in single-channel ECG data streams showed that the robustness of the results was clearly too low, so that this path could not be pursued further.

In a second, alternative, recurrent-network-based approach ("Pose+RNN"), the pre-processing was replaced by the ESN, using instead the generation of pose-frame data streams from the baseline approach, and a downstream activity detection from pose-frame data streams using LSTM was used to solve the task, which gave very good results and served as a subsequent benchmark to the use of SNNs on data streams from the event camera.

M3.2 Training and evaluation of activity recognition from frame camera

Data streams: Prior to the collection of project-specific data, the "Pose+t-CNN" processing network was trained and evaluated with data from the publicly available NTU RGB-d dataset for the detection of selected activities, which demonstrated state-of-the-art performance.

M3.3 Training and evaluation of activity recognition from event frame-

Data streams: Instead of the originally envisaged network for processing event frame data streams, which proved unsatisfactory in the preliminary investigations during the development of the architecture, "Pose+RNN" was developed as an alternative, as explained in M3.1. Similarly, the "Pose+RNN" network was trained on pose frame data obtained from the NTU RGB-d dataset and found to be roughly equivalent to the "Pose+t-CNN" processing network before performance. The "Pose+RNN" network was then selected to compete against a SNN (which processes data directly from the event camera) and trained accordingly using the data from the Dante set collected as part of the project.

M4.1 Selection of Spiking Neural Networks: The project initially investigated various frameworks for the development of Spiking Neural Networks (SNN), as these have a direct influence on the possible SNN models. Initial tests with Nengo[1], PyNN[2] and BindsNET[3] were discontinued due to lack of functionality or complex extensibility/integration. In the end, Norse[4] in combination with Tonic[5] was used as a framework because they promised sufficient functionality, fast results and easy integration of event camera data. Recurrent, convolutional and fully networked SNN models were chosen.

M4.2 Training of Spiking Neural Networks: Several SNN models were trained on different datasets. Initially, synthetic data based on the conversion algorithm developed for M2.2 was used. Later, a dataset consisting of twelve classes and a total of 580 images, each recorded with one frame camera and two event cameras in parallel, was included in the project. Of these twelve classes, ten classes are based on the neuromorphic "DVS Gesture" dataset[6] , which are extended by the classes "DVS Gesture" and "DVS Gesture".

"Sit down" and "Stand up" were added to account for the project context. This dataset was used to train multiple SNNs, with all models trained for 3 classes (stand up, sit down, other) and 12 classes (all of the above).

M4.3 Evaluation of the spiking neural networks: The trained SNNs were evaluated based on the recorded data, especially with regard to prediction accuracy and number of parameters. In particular, the number of parameters should not be too large in order to ensure usability in embedded systems. As a guideline, ~ 5 million parameters were chosen. The evaluation showed that pure fully wired and recursive models could not deliver competitive prediction accuracy without greatly exceeding the target number of parameters. Better results were obtained with convolutional nets, which have a

---

[1] https://www.nengo.ai/
[2] https://neuralensemble.org/PyNN/
[3] https://github.com/BindsNET/bindsnet
[4] https://github.com/norse/norse
[5] https://github.com/neuromorphs/tonic
[6] https://research.ibm.com/interactive/dvsgesture/

achieve similar accuracy as KNN developed to M3.2. The best results were obtained with a combination of recurrent, fully wired and convolutional layers, which achieved a very high accuracy of about 98% for 3 classes with less than
400,000 parameters. A detailed list and comparison of the different networks and accuracies can be found in Table 2.

Table 2: Comparison of accuracy of the grids on the recorded data

Neural network	Accuracy for 3 classes	Accuracy for 12 classes	Number of parameters
ANN (CNN + LSTM)	96%	57%	~ 6.500.000
SNN Fully Connected	80%	30%	~ 10.000.000
SNN Convolutional	95%	87%	~ 1.000.000
SNN Conv. + Recurrent	98%	92%	~ 380.000

The SNNs were also ported as far as possible to the neuromorphic hardware platform BrainChip Akida[7] and evaluated there in terms of throughput and electrical power consumption. A comparison with the simulation of the SNN on a GPU-based computer and with the values of the KNN-based system can be seen in Table 3.

Table 3: Comparison of el. power and BPS of the neural networks on different hardware systems

System	Power consumption	Images per second
KNN on Nvidia Jetson NX	~5-9W	~1
SNN on GPU computer	~150W	0,5 - 1
SNN on BrainChip Akida	<1W	40 - 45

M5.1 Evaluation and demonstration of the project results: To demonstrate the project results, a system was set up that integrates one classical and one event camera. In addition, an Nvidia Jetson NX[8] as a classical AI processor and a BrainChip Akida as a neuromorphic AI processor are connected to the system, running the neural networks that the partners developed within the project. The videos from both cameras are fed to a graphical

[7] https://brainchipinc.com/akida-neural-processor-soc/

[8] https://www.nvidia.com/de-de/autonomous-machines/embedded-systems/jetson-xavier-nx/

user interface so that an observer can see the differences between the various cameras. In addition, the data from the event camera is forwarded to the neuromorphic processor and processed there by an SNN, while the data from the classic camera is forwarded to the Nvidia Jetson NX. The concept of the demonstrator is shown in Figure 4.

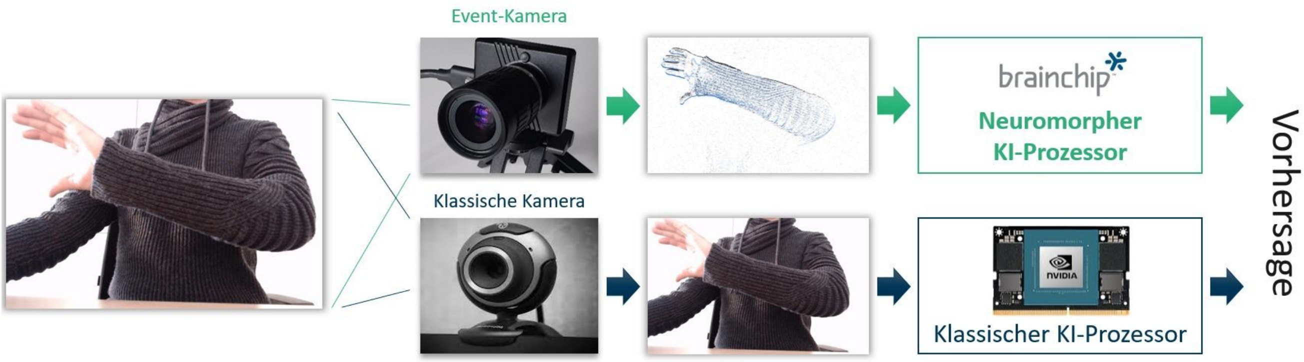


Figure 4: Conceptual structure of the demonstrator

The predictions of the respective networks as well as the required power consumption and the data rate currently transmitted by the cameras are also graphically displayed in parallel. The graphical interface can be seen in Figure 5.

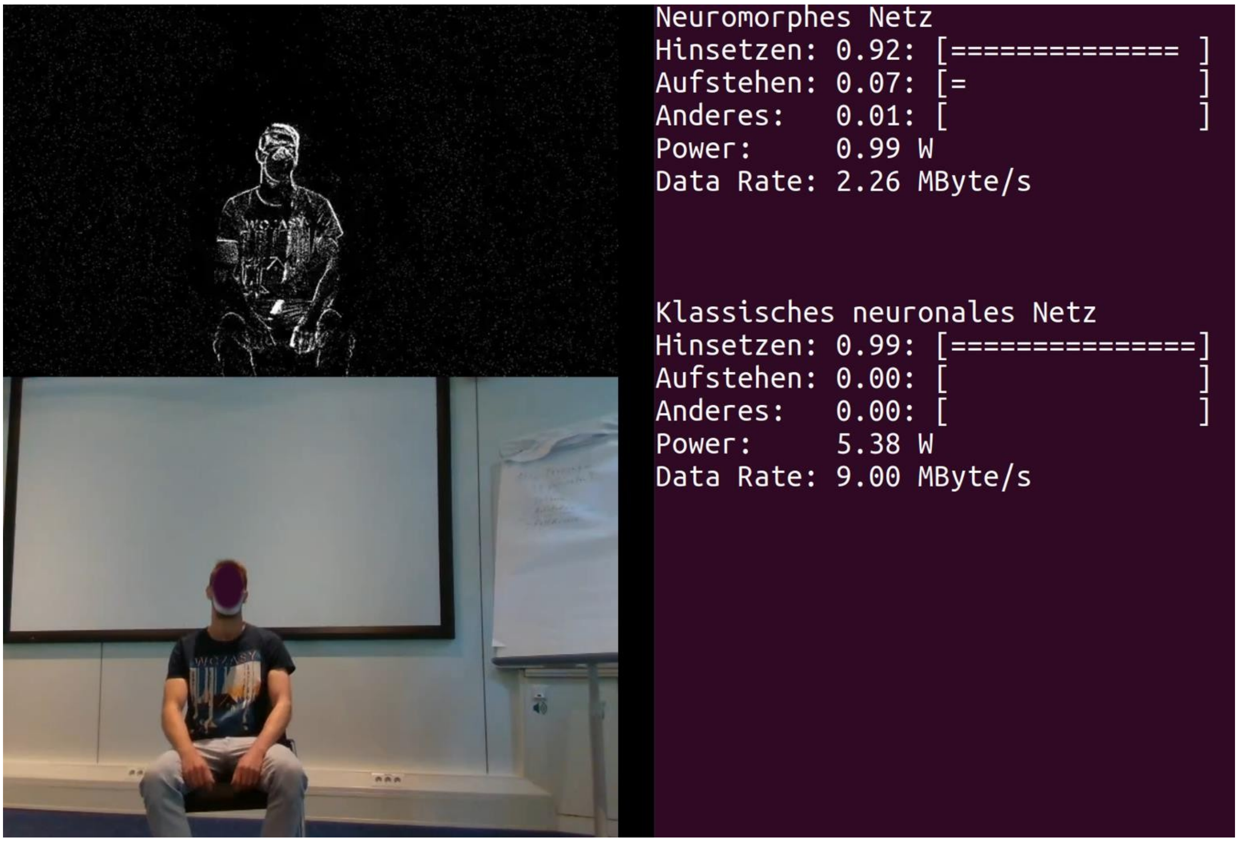


Figure 5: Recording of the demonstrator interface during operation

The demonstration shows the potential of the neuromorphic system in terms of energy efficiency and the amount of data required, both of which have been improved by a factor of about 5. However, this is offset by the high costs of the neuromorphic system and in particular the event camera of several thousand euros, which still have to be significantly reduced before they can be successfully used in the application studied.

Deviation from the original planning

Hardware procurement

Due to the delay in the delivery of the camera from China and a very high effort on the administrative side or the defective delivery of the hardware from France and renewed deliveries, there was a delay in the data creation and subsequent integration from the data format into a format necessary for the project. Under time pressure, it was even possible to solve some technical problems during the integration, but this was done at a higher cost.

effort than planned on the resource side. In the end, we could not leave ourselves enough time for the evaluation of different SNN architectures and limited the investigation to one version of an SNN architecture and its implementation, which depicts the basic features of the project, but could be optimised in further projects. The creation of the application and the integration into software took place without problems and according to the previously defined plan.

SNN architecture

As already described in M3.1 and M3.3, the originally planned neural network architecture for activity recognition by means of recurrent neural networks using echo state networks for processing event frame data streams did not prove successful. Therefore, a new architecture was developed, which uses LSTM as a recurrent network for the recognition of activities, but is based on pose data streams.

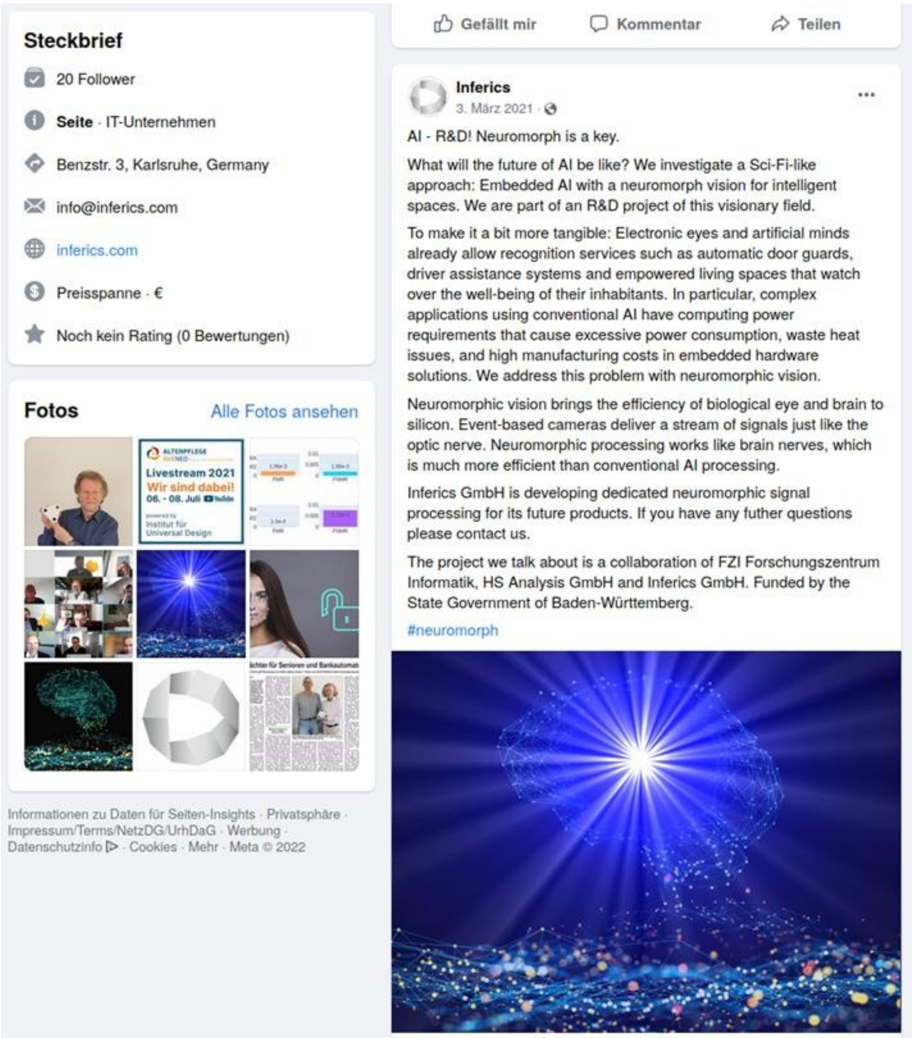
2. Dissemination of the results

The project and its results were presented to the public by all partners on many different channels, including conferences, festivals, lectures, websites, social media and other events. The following is a detailed list:

- Presentation of the ongoing project at the science festival "Effects", which was organised by the city of Karlsruhe and the Karlsruhe Institute of Technology for the general public in summer 2021. https://www.effekte-karlsruhe.de/festival-2021/programm-archiv/
- The ongoing project was presented at the Artificial Intelligence Working Group of the Karlsruhe Chamber of Industry and Commerce (IHK). The focus here was on the cooperation between FZI, Inferics GmbH and HS Analysis GmbH as experts for SNNs.
- At the annual meeting of the Karlsruhe Chamber of Industry and Commerce on the premises of HS Analysis GmbH, HS Analysis was able to give a practical live presentation of progress in the use of the event camera and SNNs. Participants of this

The discussion panel included the company IONOS, as well as the Ministry for Research and Economy of the State of Baden-Württemberg.

- Presentation on 29.09.2021 at the KI4KMU track of the conference INFORMATIK2021, hosted by the Gesellschaft für Informatik, in combination with a publication.
- Presentation at the Tollhaus Karlsruhe in November 06.10.2021 KIT Science Week.
- Posts on Facebook
(https://www.facebook.com/people/Inferics/100057143332472/)
and LinkedIn
(https://de.linkedin.com/company/inferics).



- Article on https://www.inferics.com/news/
- Presentation to partners from science, industry and politics at the FZI OpenHouse 2022 on 24.02.2022:

https://www.fzi.de/veranstaltungen/fzi-open-house/ - 12:15 & 13:15 | Neuromorphic Systems and Sensors

3. Exploitation and transfer

Knowledge transfer

The collaboration in the project was characterised by a strong knowledge transfer between the partners, as each partner brings its own core competence to the project. HSA focused on the general analysis of data with different deep learning architectures. Thus, by working in consortium with Inferics, it is possible to explore and apply the data to the use case with new deep learning technologies. The focus of Inferics was on sensors for the recognition of activities, which are to enable battery-powered interactive assistance for people over a period of weeks. Inferics provided the application reference and developed simulated neuromorphic structures, which are contrasted with the "real" neuromorphic systems of the partners HSA and FZI. This led to an intensive exchange of experiences on capabilities and limitations as well as on the necessary efforts and savings of the competing approaches. FZI contributed its extensive knowledge of neuromorphic algorithms and sensor technology to support both partners in adapting their technologies to a common neuromorphic system.

The mutual exchange of knowledge resulted in a common understanding of the technologies used, which enabled the project partners to build a meaningful demonstrator and successfully complete the project.

Transfer concept for commercial exploitation

HS Analysis

HS Analysis has experienced great interest from our existing customers as well as some potential new customers to apply SNNs in commercial products, especially in the area of process monitoring. The main interest here lies in the reduced energy consumption and increased data protection of event-based cameras. This enables data protection-compliant process monitoring even in areas accessible to the public, and the reduced power requirement offers the possibility of operating on-edge devices via Power-over-Ethernet (PoE), which enables simplified deployment.

We can also extend our previous core product, the HSA KIT, which contains a toolbox of various customised AI analyses, with the SNN knowledge gained in the project. This means that advanced, minimalistic time series analyses are now possible, for which we have already started the process of quoting and ordering.

Inferics

Whether patient monitoring in hospitals or emergency and activity detection in assisted living or nursing homes - the assumption of routine nursing tasks counters the much-cited nursing shortage and therefore has enormous economic potential, which Inferics GmbH opens up through appropriate devices for recording and reporting. In order to guarantee the privacy of care recipients, processing takes place in the sensor itself, so that only messages leave the sensor. However, this also means having to perform complex recognition services in sensors with a small form factor. With conventional camera and neural network technology, this leads to severe limitations. In the project it became apparent that these can be overcome by means of event cameras and processing by much lighter SNN (especially with the availability of hardware that is specifically efficient for this purpose). At present, however, the very high price of the hardware (a multiple of the achievable total system costs) clearly speaks against the marketability of such solutions. However, it is worthwhile for Inferics to observe the price development in order to be able to switch to the new technology in time based on the experience gained in the project. As an interim result of the project, it was recognised that recurrent networks can advantageously replace the time convolution networks used up to now due to their lower resource requirements, so that several detection performances are possible on one sensor at the same time, which represents a clear market advantage.

can result. In the project, SNNs proved to be even more resource-efficient; their application to event frame data (generated from frame data of a conventional frame camera) can multiply the effect described. The latter could result in a real breakthrough in terms of price, which would significantly increase the market potential.