TinyML Journey - contextual awareness for laptop PCs

Elad Baram, VP Products, eladb@emza-vs.com

17 March 2021
About emza

• Founded in 2006 in Israel, with a vision for low power edge computing

• Develops & sells ultra-low power computer vision solutions (HW & SW)

• Acquired by Himax Technologies in 2018. (NASDAQ: HIMX)
The starting point

CES 2019, Intel announced project Athena

Opportunity:

provide contextual understanding in privacy, at low power based on Ai
AI sensing landscape in laptop devices 2019

- **Shapes**
  - Proximity sensor (ToF)
  - Gyroscope & accelerometer
- **Objects, humans**
  - Human detection
  - Face detection
  - Face orientation
  - Face recognition
  - Face expression
  - Eye tracking
- **Identities, emotions**
  - HD vision

**Sensing power**
- 1000mW
- 100mW
- 10mW
- 1mW

**Compute power**
- GHz
- MHz

**Intelligence**
- Light
- Motion
- Presence

**Private domain**
- Non-private domain

**Non-private domain**
- SW solutions

**Physical**
- ALS
- Light
- Motion
- Presence
AI landscape with visual sensing

- **Shapes**
  - Light
  - Motion
  - Presence

- **Objects, humans**
  - Human detection
  - Face detection
  - Face orientation
  - Face recognition
  - Face expression
  - Eye tracking

- **Identities, emotions**
  - Visual sensing
  - Emza
  - Proximity sensor (ToF)
  - Gyroscope & accelerometer

- **Physical**
  - ALS
  - Compute power
    - GHz
  - Non-private domain
  - SW solutions
  - HD vision
  - Private domain

- **Classification**
  - Sensing power
    - mW
    - 10mW
    - 100mW
    - 1000mW
  - Private domain
  - Non-private domain

- **Intelligence**
Visual sensing in PC – system architecture

- Human presence detection algorithms
- WiseEye1 ULP CV ASIC
- Integrated within the camera module
- Always-on sensing
- User privacy guaranteed
# Human Presence Detection (HPD)

## Use cases

<table>
<thead>
<tr>
<th><strong>Wake on approach</strong></th>
<th><strong>User presence</strong></th>
<th><strong>Walk away lock</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic wake up</td>
<td>Classification – humans vs objects</td>
<td>Automatic screen lock – security</td>
</tr>
<tr>
<td>Touchless Hello/Login experience</td>
<td></td>
<td>Automatic screen off – power save</td>
</tr>
</tbody>
</table>
CES 2020 demo: Wake on Approach
Contextual privacy

Understanding the context and actively improve privacy security

Sensing

Intruder alert

System response

Auto screen blur

Auto screen dim
User engagement detection at low power

Opportunity to extend battery life by 20%
• New concept: user engagement status
• Engaged: frontal face detected
• Not engaged: user is not looking on the display

New classification requirements:
• Yaw angle
• Head classifier

extends battery life with Adaptive dimming

Engaged
- high illumination

Not Engaged
- dimming
Real World Challenges

• The Distance / Field of view / execution speed tradeoff

• Faces in the wild

• How can one detect the engagement level?

• Hard illumination condition as the typical environment
Distance / Field of view / speed challenge

Requirements:

Distance: 25-200cm
Horizontal field of view: 70°
Processing: 8-10 FPS
Output: bounding boxes
Moving to Detector

Motivation for detector

• Bounding box (object location)
• User distance estimation (based on face size)
• Tracking (save power)
• Advanced classification options

But,

running SSD on microcontroller in a naïve manner is merely impossible...
Distance / Field of view / FPS challenge

96x96 MobileNet TFLM

FPS

1MB, mW scale

FoV

Distance

Cropping with Mobile net

Tiling with Mobile net
“micro” CV detector pipeline

- Classical ML for fast box proposal
- Deep learning for short distance / within the boxes

Stage 1: Detection
- Fast face detector with Classical ML

Stage 2: Classification
- Advanced classifiers within the proposed boxes
Putting it all together

Face detection & yaw angle estimation video
And then Covid-19...
Real world variety

Natural user posing
Real world variety

It is not all about Lux, it’s the dynamic range
TinyML in reality
Ai enabled camera module for laptops
The journey has just begun

- TinyML brings value in consumers applications
- We are in the early days of adoption
- Innovation in algorithms & silicon IP will enable more sophisticated use cases and will accelerate adoption
Thank you
We thank the authors for their presentations and everyone who participated in the tinyML Summit 2021.

Along with a special thank you to the sponsors who made this event possible!
Executive Sponsors
Arm: The Software and Hardware Foundation for tinyML

Connect to high-level frameworks

1. Profiling and debugging tooling such as Arm Keil MDK

2. Supported by end-to-end tooling

3. Connect to Runtime

Application

Optimized models for embedded

Runtime (e.g. TensorFlow Lite Micro)

Optimized low-level NN libraries (i.e. CMSIS-NN)

RTOS such as Mbed OS

Arm Cortex-M CPUs and microNPUs

AI Ecosystem Partners

Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm
Advancing AI research to make efficient AI ubiquitous

Perception
Object detection, speech recognition, contextual fusion

Reasoning
Scene understanding, language understanding, behavior prediction

Action
Reinforcement learning for decision making

Power efficiency
Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization
Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning
Robust learning through minimal data, unsupervised learning, on-device learning

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
Samsung brings AI in the hands of everyone, with >300M Galaxy phones per year. Fingerprint ID, speech recognition, voice assistant, machine translation, face recognition, AI camera; the application list goes on and on.

In the heart of AI applications is the NPU, the neural processor that efficiently calculates AI workloads. Samsung NPU is a home grown IP that was employed since 2018 inside Samsung Exynos SoC.

Samsung NPU is brought by global R&D ecosystem that encompasses US, Korea, Russia, India, and China. In US, we are the fore-runner to guide the future directions of Samsung NPU, by identifying major AI workloads that Samsung’s NPU needs to accelerate in 3-5 years. For this, we collaborate with world-renowned academia research groups in AI and NPU.
Platinum Sponsors
**Eta Compute** creates energy-efficient AI endpoint solutions that enable sensing devices to make autonomous decisions in energy-constrained environments in smart infrastructure and buildings, consumer, medical, retail, and a diverse range of IoT applications.

[www.etacompute.com](http://www.etacompute.com)
Lattice Semiconductor (NASDAQ: LSCC) is the low power programmable leader. We solve customer problems across the network, from the Edge to the Cloud, in the growing communications, computing, industrial, automotive and consumer markets. Our technology, relationships, and commitment to support lets our customers unleash their innovation to create a smart, secure and connected world. www.Latticesemi.com.
Gold Sponsors
AKIDA™ Neuromorphic Technology: Inspired by the Spiking Nature of the Human Brain

- Supports ultra-low power applications (microwatts to milliwatts)
- Edge capabilities: on-chip training, learning, and inference
- Designed for AI Edge applications: vision, audio, olfactory, and smart transducer applications
- Licensed as IP to be designed into SoC or as silicon
- Sensor inputs are analyzed at the point of acquisition rather than through transmission via the cloud to the data center. Enables real-time response for power-efficient systems
- Software Development Platform
BabbleLabs AI speech wizardry in Cisco Webex

AI meets speech - deep experience in speech science, AI/ML, embedded systems

Massive compute

- Novel deep neural networks

Massive data corpus

- 300 TFLOPS per engineer
- Silicon-optimized software
- 40K hours of speech
- 15K hours of music
- 10K hour of noise
- 100K room models

Speech enhancement

- Speech recognition

Conferencing

- Call centers
- Digital Assistants
- Calling
DSP Group, Inc. develops wireless communications and voice processing chipsets, algorithms, and software solutions for converged communications and smart-enabled devices. Core competencies include, but are not limited to, voice processing. Its technology supports the development and integration of voice user interfaces (VUIs) for applications ranging from smartphones to the smart home. Its Ultra-Low Energy (ULE, per the ULE Alliance) wireless solutions enable low-power, long-range, secure communication applications for the IoT and are distinguished by their native support of two-way voice communication. On-going development efforts include the application of machine learning (ML) and artificial intelligence (AI) hardware and algorithms to address the need for accurate AI solutions at the edge for applications such as sound detection, proximity detection, and acoustic beacons.
TinyML for all developers

- **Dataset**: Acquire valuable training data securely
- **Impulse**: Enrich data and train ML algorithms
- **Test**: Test impulse with real-time device data flows
- **Edge Device**: Real sensors in real time
  - Open source SDK
- **Embedded and edge compute deployment options**: Enrich data and train ML algorithms

**TinyML for all developers**

- **C++ library**
- **Arduino library**
- **WebAssembly**

**www.edgeimpulse.com**
The Eye in IoT
Edge AI Visual Sensors

CMOS Imaging Sensor
- Ultra Low power CMOS imager
- AI + IR capable

IoT System on Chip
- Machine Learning edge computing silicon
- <1mW always-on power consumption
- Computer Vision hardware accelerators

Computer Vision Algorithms
- Machine Learning algorithm
- <1MB memory footprint
- Microcontrollers computing power
- Trained algorithm
- Processing of low-res images
- Human detection and other classifiers

info@emza-vs.com
GrAI Matter Labs has created an AI Processor for use in edge devices like drones, robots, surveillance cameras, and more that require real-time intelligent response at low power. Inspired by the biological brain, its computing architecture utilizes sparsity to enable a design which scales from tiny to large-scale machine learning applications.

www.graimatterlabs.ai
Enabling the next generation of **Sensor and Hearable products**
to process rich data with energy efficiency

- Visible Image
- Sound
- IR Image
- Radar
- Bio-sensor
- Gyro/Accel

Wearables / Hearables
Battery-powered consumer electronics
IoT Sensors
Himax Technologies, Inc. provides semiconductor solutions specialized in computer vision. Himax’s WE-I Plus, an AI accelerator-embedded ASIC platform for ultra-low power applications, is designed to deploy CNN-based machine learning (ML) models on battery-powered AIoT devices. These end-point AI platforms can be always watching, always sensing, and always listening with on-device event recognition.

Imagimob AI SaaS

• End-to-end development of tinyML applications
• Guides and empowers users through the process
• Support for high accuracy applications requiring low power and small memory
• Imagimob AI have been used in 25+ tinyML customer projects
• Gesture control
Adaptive AI for the Intelligent Edge

LatentAI

Latentai.com
<table>
<thead>
<tr>
<th>Advanced AI Acceleration IC</th>
<th>Low Power Cortex M4 Micros</th>
<th>Sensors and Signal Conditioning</th>
</tr>
</thead>
<tbody>
<tr>
<td>The new MAX78000 implements AI inferences at low energy levels, enabling complex audio and video inferencing to run on small batteries. Now the edge can see and hear like never before.</td>
<td>Large (3MB flash + 1MB SRAM) and small (256KB flash + 96KB SRAM, 1.6mm x 1.6mm) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels.</td>
<td>Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.</td>
</tr>
</tbody>
</table>

www.maximintegrated.com/MAX78000  
www.maximintegrated.com/microcontrollers  
www.maximintegrated.com/sensors
Qeexo AutoML

Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

Key Features

- Supports 17 ML methods:
  - Multi-class algorithms: GBM, XGBoost, Random Forest, Logistic Regression, Gaussian Naive Bayes, Decision Tree, Polynomial SVM, RBF SVM, SVM, CNN, RNN, CRNN, ANN
  - Single-class algorithms: Local Outlier Factor, One Class SVM, One Class Random Forest, Isolation Forest
- Labels, records, validates, and visualizes time-series sensor data
- On-device inference optimized for low latency, low power consumption, and small memory footprint applications
- Supports Arm® Cortex™- M0 to M4 class MCUs

End-to-End Machine Learning Platform

For more information, visit: www.qeexo.com

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT
Add Advanced Sensing to your Product with Edge AI / TinyML

Pre-built Edge AI sensing modules, plus tools to build your own

Reality AI solutions
- Prebuilt sound recognition models for indoor and outdoor use cases
- Solution for industrial anomaly detection
- Pre-built automotive solution that lets cars “see with sound”

Reality AI Tools® software
- Build prototypes, then turn them into real products
- Explain ML models and relate the function to the physics
- Optimize the hardware, including sensor selection and placement

https://reality.ai    info@reality.ai    @SensorAI    Reality AI
Build Smart IoT Sensor Devices From Data

SensiML pioneered TinyML software tools that auto generate AI code for the intelligent edge.

- End-to-end AI workflow
- Multi-user auto-labeling of time-series data
- Code transparency and customization at each step in the pipeline

We enable the creation of production-grade smart sensor devices.

sensiml.com
Silicon Labs (NASDAQ: SLAB) provides silicon, software and solutions for a smarter, more connected world. Our technologies are shaping the future of the Internet of Things, Internet infrastructure, industrial automation, consumer and automotive markets. Our engineering team creates products focused on performance, energy savings, connectivity, and simplicity. silabs.com
Syntiant Corp. is moving artificial intelligence and machine learning from the cloud to edge devices. Syntiant’s chip solutions merge deep learning with semiconductor design to produce ultra-low-power, high performance, deep neural network processors. These network processors enable always-on applications in battery-powered devices, such as smartphones, smart speakers, earbuds, hearing aids, and laptops. Syntiant's Neural Decision Processors™ offer wake word, command word, and event detection in a chip for always-on voice and sensor applications.

Founded in 2017 and headquartered in Irvine, California, the company is backed by Amazon, Applied Materials, Atlantic Bridge Capital, Bosch, Intel Capital, Microsoft, Motorola, and others. Syntiant was recently named a CES® 2021 Best of Innovation Awards Honoree, shipped over 10M units worldwide, and unveiled the NDP120 part of the NDP10x family of inference engines for low-power applications.
TensorFlow is an end-to-end open source platform for machine learning. Our ecosystem of tools, libraries, and community resources help users push the state-of-the-art in building and deploying ML powered applications.
JOIN OUR SESSIONS DURING THE TINYML SUMMIT

Performing inference on BNNs with xcore.ai
Tuesday, March 23 at 12pm (PST)

TinyML: The power/cost conundrum
Thursday, March 25 at 12pm (PST)

VISIT XMOS.AI TO FIND OUT MORE
Silver Sponsors

EDGE
CORTIX
HOTG
SynSense
Copyright Notice

The presentation(s) in this publication comprise the proceedings of tinyML® Summit 2021. The content reflects the opinion of the authors and their respective companies. This version of the presentation may differ from the version that was presented at the tinyML Summit. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyML.org