symmetry

Article

Smart Doll: Emotion Recognition Using Embedded Deep Learning

Jose Luis Espinosa-Aranda [1], Noelia Vallez [1], Jose Maria Rico-Saavedra [1], Javier Parra-Patino [1], Gloria Bueno [1], Matteo Sorci [2], David Moloney [3], Dexmont Pena [3] and Oscar Deniz [1,*]

[1] VISILAB, University of Castilla-La Mancha, E.T.S.I. Industriales, Avda Camilo Jose Cela s/n, 13071 Ciudad Real Spain; josel.espinosa@uclm.es (J.L.E.-A.); noelia.vallez@uclm.es (N.V.) josemaria.rico@uclm.es (J.M.R.-S.); javier.parra@uclm.es (J.P.-P.); gloria.bueno@uclm.es (G.B.)

[2] nViso SA, PSE-D, Site EPFL, CH-1015 Lausanne, Switzerland; matteo.sorci@nviso.ch

[3] Intel R&D Ireland Ltd., Collinstown Industrial Park, Leixlip, Co Kildare W23 CW68, Ireland; david.moloney@intel.com (D.M.); dexmont.pena@intel.com (D.P.)

[*] Correspondence: JoseL.Espinosa@uclm.es or Oscar.Deniz@uclm.es

Received: 13 July 2018; Accepted: 7 September 2018; Published: 7 September 2018

Abstract: Computer vision and deep learning are clearly demonstrating a capability to create engaging cognitive applications and services. However, these applications have been mostly confined to powerful Graphic Processing Units (GPUs) or the cloud due to their demanding computational requirements. Cloud processing has obvious bandwidth, energy consumption and privacy issues. The Eyes of Things (EoT) is a powerful and versatile embedded computer vision platform which allows the user to develop artificial vision and deep learning applications that analyse images locally. In this article, we use the deep learning capabilities of an EoT device for a real-life facial informatics application: a doll capable of recognizing emotions, using deep learning techniques, and acting accordingly. The main impact and significance of the presented application is in showing that a toy can now do advanced processing locally, without the need of further computation in the cloud, thus reducing latency and removing most of the ethical issues involved. Finally, the performance of the convolutional neural network developed for that purpose is studied and a pilot was conducted on a panel of 12 children aged between four and ten years old to test the doll.

Keywords: facial informatics; deep learning; computer vision; mobile applications; real-time and embedded systems

1. Introduction

Traditionally focused on factory automation, computer vision (i.e., software that automatically analyses images to extract content and meaning) is rapidly extending several new scenarios. Vision technology allows inferring information from reality and enables new types of interactivity. There are several examples of successful computer vision devices and applications, such as Microsoft's Kinect, Google Glass, Microsoft HoloLens, Dyson's vacuum cleaners, Amazon Fire, etc., which have been generally developed by large companies able to afford the resources needed for the application-specific development. While some required technologies are already mature and affordable, the fact is that no flexible open platform for mobile embedded vision is currently available that can be used in the context of an unprecedented rate of new mobile and ubiquitous applications.

An alternative is to use IoT (Internet of Things) devices and perform computer vision operations in the cloud [1–3]. However, the particular case of computer vision represents a fundamental challenge. In general, raw visual information cannot be transferred to the cloud for processing due to bandwidth, power consumption and security issues.

The Eyes of Things (EoT) [4] platform is an optimized core vision unit designed by the authors, that can work independently and also embedded into all types of artefacts (Figure 1) at minimal power consumption.

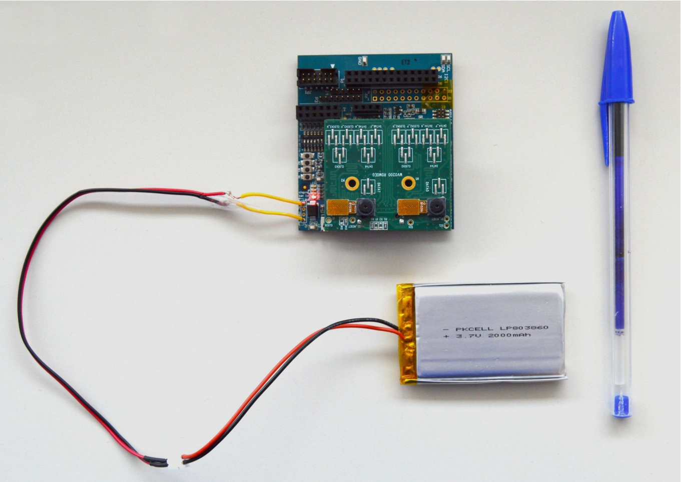


Figure 1. EoT device connected to a 3.7 V 2000 mAh flat battery.

The EoT platform is based on the Myriad 2 chip by Movidius [5,6]. Myriad 2 is a system-on-chip that embeds a software-controlled multicore, a multiported memory subsystem, and caches that can be configured to allow a large range of workloads. It provides exceptionally high sustainable on-chip data and instruction bandwidth to support high-performance video hardware accelerator filters, two main Reduced Instruction Set Computer (RISC) processors and 12 specific proprietary processors for managing computer vision and deep learning operations denominated SHAVE (Streaming Hybrid Architecture Vector Engine), a hybrid stream processor architecture designed to maximize performance-per-watt combining features of Digital Signal Processors (DSPs), GPUs, and RISC with both 8/16/32 bit integer and 16/32 bit floating point arithmetic as well as other features such as hardware support for sparse data structures. The Myriad 2 platform represents an optimization of processing power, energy consumption, size and cost [7]. EoT provides libraries and protocols which have been carefully selected and optimized for low power.

The main objective of this paper is to describe the development, using EoT of a smart doll able to interact with children. This paper is organized as follows. Section 2 explains the deep learning capabilities of the EoT device. Section 3 shows and validates the convolutional neural network for emotion recognition developed and Section 4 analyses its computational performance. Section 5 presents the smart doll developed and Section 6 the pilot conducted using the doll. Finally, Section 7 outlines the conclusions of this work.

2. Deep Learning in EoT

Although several computational algorithms and methods have shown their performance and capacity [8–10], the deep learning paradigm no doubt represents a breakthrough in machine learning and consequently in computer vision [11,12]. Both academia and industry are currently moving at lightspeed towards this methodology [13]. While EoT provides multiple software libraries and applications for computer vision and communication with the outside world, deep learning is obviously one of the key elements. The deep learning capabilities in EoT, specifically in the form of convolutional neural networks (CNN), are contained in the following modules:

2.1. Tiny_dnn

This library [14] is an open source C++ deep learning inference engine suitable for limited computational resources, embedded systems and IoT devices. It includes several features that makes it perfect for EoT: it is a header-only library with no external dependences and it can import Caffe models. Moreover, it has been recently included as the deep learning engine of OpenCV since version 3.1, evidencing its increasing importance.

The library was streamlined and ported to EoT. This version allows reading a network stored in tiny_dnn's JavaScript Object Notation (JSON) format which requires training the network from scratch using the PC version of the library or converting it from Caffe's format (Figure 2).

Moreover, some of the most time-consuming parts of this library were parallelised by making use of the 12 SHAVE processors of the Myriad SoC.

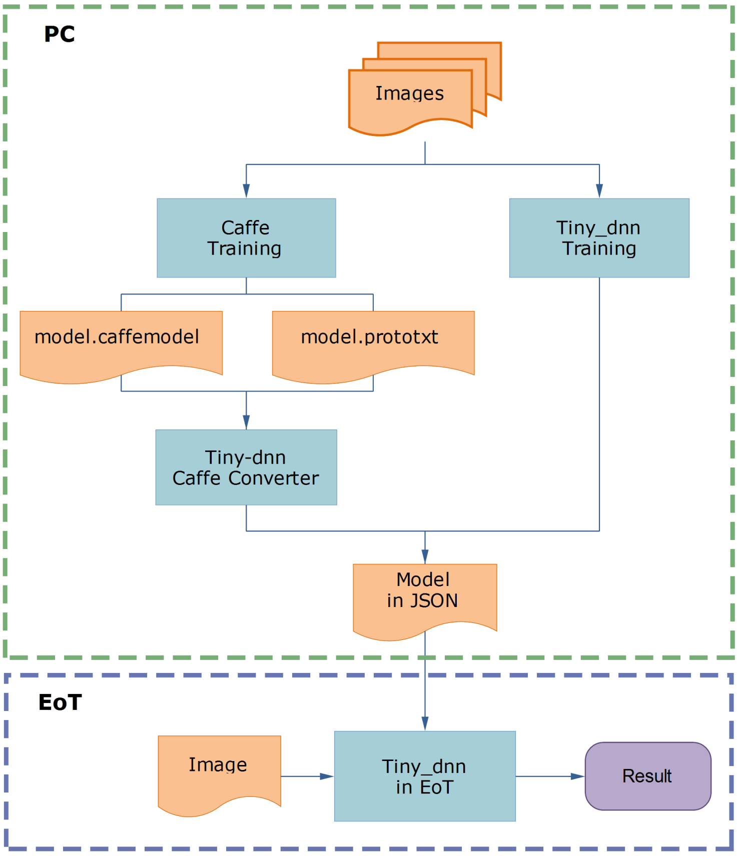


Figure 2. Tiny_dnn workflow.

2.2. Fathom Framework

In order to address the deep learning paradigm, Movidius developed a proprietary framework called Fathom to execute CNNs of various configurations at ultra-low power targeting the Myriad 2 SoC. The Fathom framework accepts eXtensible Markup Language (XML) representations of trained networks from Caffe or TensorFlow. Fathom parses and optimally transforms them to run on the Myriad 2 architecture. For complex networks such as GoogleNet, Fathom enables performance at a nominal 15 inferences/sec with fp16 precision on Myriad 2.

Although Fathom offers a significant improvement in terms of performance, it has the drawback that it was originally developed to be used with the Movidius$^{\text{TM}}$Neural Compute Stick [15]. This was not useful in a standalone platform such as EoT. We have been able to port Fathom to EoT by establishing some limitations while modifying the original library. While the Fathom framework expects an input image stored as a file, we changed this so as to get the image directly from the camera of the EoT device using the specific input/output of the platform. Moreover, the modified framework also allows to load the trained network directly from the microSD card included in EoT whereas the original framework received it from the PC to which it was connected (through USB). In this case, it is necessary to store the network in the format used by the framework, a so-called blob file, which can be generated from a Caffe model using the tools provided by the original Fathom framework on a PC. Finally, some modifications were made with respect to cache management in the EoT platform to solve coherency problems arising with the execution of consecutive inferences.

3. Convolutional Neural Network for Emotion Recognition

3.1. Description of the Emotion Recognition Model

Identifying human emotions is not a simple task for machines. As human beings this comes natural to us, but traditional machine learning methodologies struggle to achieve satisfactory results. For machines to predict emotions by analysing the facial expressions, it is necessary to identify the basic set of emotions which are universal and can be quantified.

Ekman and associates [16,17], identified the set of six universal basic emotions: anger, disgust, fear, happiness, sadness and surprise.

The proposed nViso emotion engine is trained to classify the six basic emotions plus the neutral expression, Figure 3 provides an example of the emotion classes.

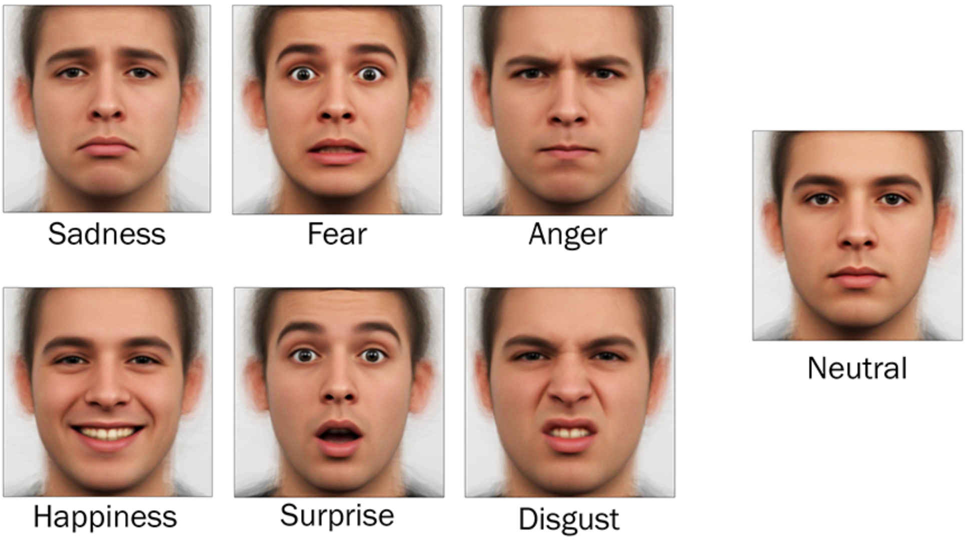


Figure 3. Emotion classes predicted by the emotion engine.

The emotion engine uses a specially designed CNN for emotion recognition. In order to produce a model small enough to run on a low powered embedded platform like EoT, the network was trained on a relatively smaller dataset comprising 6258 images obtained from 700 different subjects. The network has seven Rectified Linear Unit (RELU) layers and three pooling layers, see Figure 4.

As depicted in Figure 4, the input to network is an image of size 50 × 50 containing the face. The network applies the series of RELU and pooling layers with the output layer having seven nodes, one for each emotion. The trained network occupies a total of 900 KB of storage space and is stored on the EoT microSD card. This card can be removed from the EoT board and read/written on a PC.

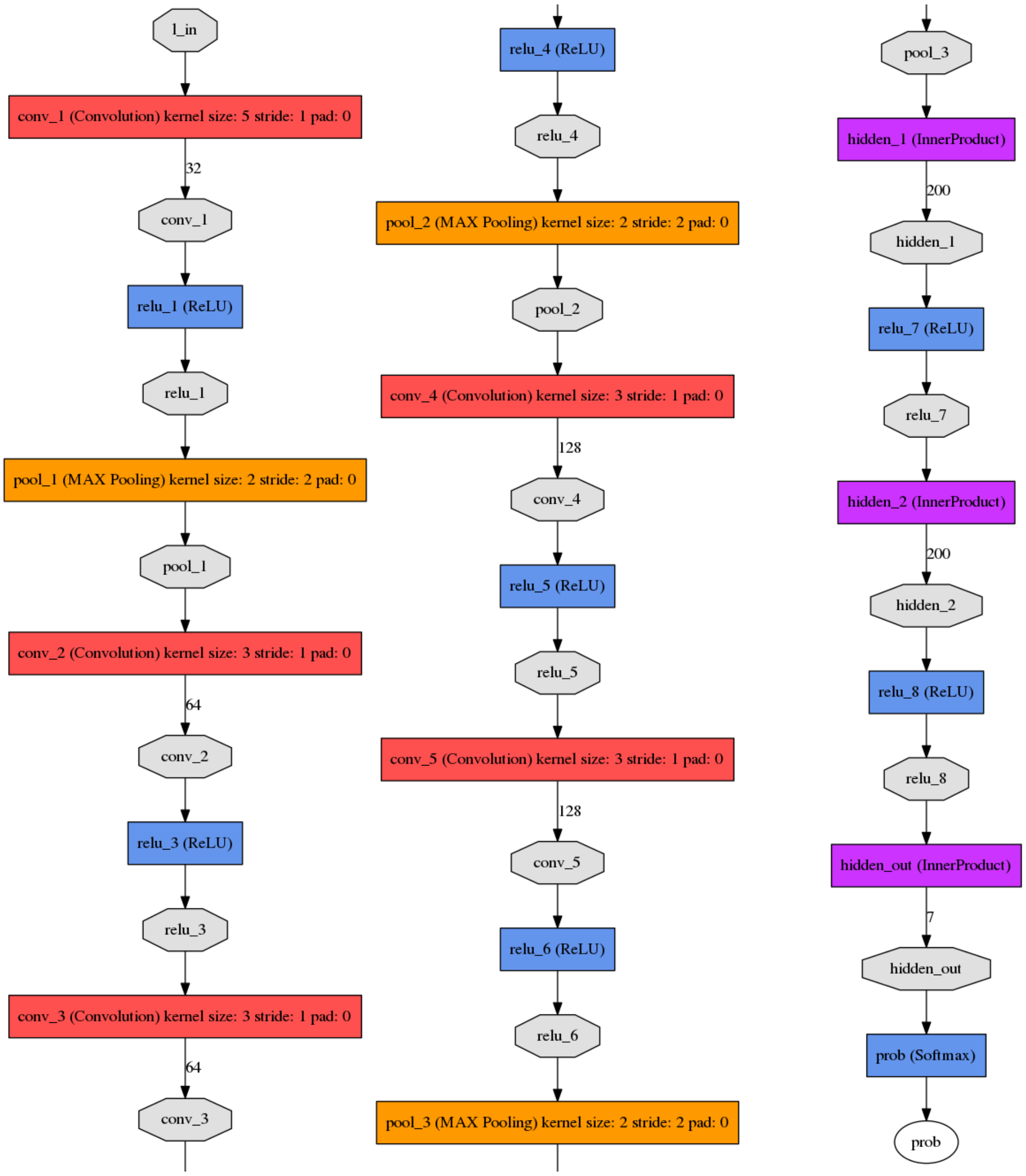

l_in
conv_1 (Convolution) kernel size: 5 stride: 1 pad: 0
32
conv_1
relu_1 (ReLU)
relu_1
pool_1 (MAX Pooling) kernel size: 2 stride: 2 pad: 0
pool_1
conv_2 (Convolution) kernel size: 3 stride: 1 pad: 0
64
conv_2
relu_3 (ReLU)
relu_3
conv_3 (Convolution) kernel size: 3 stride: 1 pad: 0
64
conv_3
relu_4 (ReLU)
relu_4
pool_2 (MAX Pooling) kernel size: 2 stride: 2 pad: 0
pool_2
conv_4 (Convolution) kernel size: 3 stride: 1 pad: 0
128
conv_4
relu_5 (ReLU)
relu_5
conv_5 (Convolution) kernel size: 3 stride: 1 pad: 0
128
conv_5
relu_6 (ReLU)
relu_6
pool_3 (MAX Pooling) kernel size: 2 stride: 2 pad: 0
pool_3
hidden_1 (InnerProduct)
200
hidden_1
relu_7 (ReLU)
relu_7
hidden_2 (InnerProduct)
200
hidden_2
relu_8 (ReLU)
relu_8
hidden_out (InnerProduct)
7
hidden_out
prob (Softmax)
prob

Figure 4. Emotion recognition network architecture.

3.2. Validation of the Emotion Model

For the purpose of validation, in this paper one of the most widely used facial expression databases has been considered, the so-called Cohn-Kanade extended database (CK+) [18], which is described in detail in the next section. The following sections describe the validation procedure, the metric used and the achieved performance by comparing them with the approach developed and deployed by Microsoft (the Oxford Project) [19].

3.2.1. Benchmarking Database: Cohn Kanade Extended

The CK+ database is the second iteration of the reference dataset in facial expression research. The database consists of 593 sequences from 23 subjects, each sequence begins with a neutral expression and proceeds to a peak expression (an emotion). Peak expressions are fully coded with the Facial Action Coding System (FACS) [20]. Figure 5 provides an example of a sequence from the CK+ dataset.

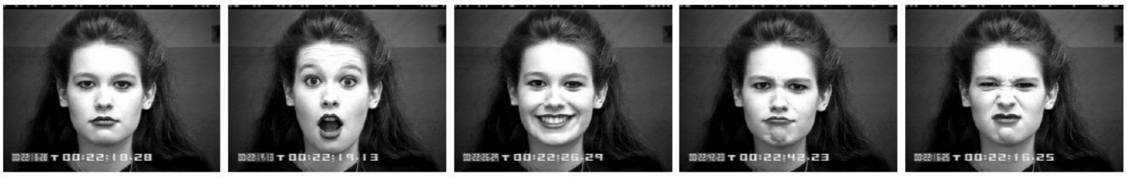


Figure 5. Example of a sequence from the Cohn-Kanade extended database (CK+) database.

FACS is a system to taxonomize human facial expressions, the concept was originally invented by Swedish anatomist Hjortsjö in 1969 [21] and it was improved and published in the following years [20]. FACS is currently the leading global standard for facial expression classification.

FACS allows to encode any facial expression, independently of any interpretation, by decomposing it into Action Units (AU). Actions units correspond to the contraction (or relaxation) of one or several muscles including one of the five levels of intensities, which vary between "Trace" and "Maximum". Moreover, every emotion is linked to a specific set of action units. Figure 6 provides some action annotations for a human face.

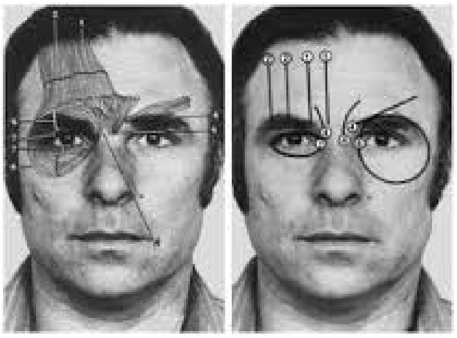


Figure 6. Action unit annotations for a human face.

Since the peak expressions in the CK+ are FACS coded, facial expressions are of very high quality, ensuring the undeniable popularity of this dataset in scientific research for the past 15 years.

3.2.2. Validation Procedure

An emotion profile is a set of normalized measures which describe the level of activation of each basic emotion. This inferred emotional state is then compared with the provided ground truth and validation metrics.

To compare the performance with state-of-the-art approaches, the Oxford model introduced by Microsoft in 2015 is used [19]. Oxford provides an online API to compute an emotion profile for a given image.

3.2.3. Validation Metrics

The Cohn-Kanade extended database consists of sequences beginning with zero-intensity expression, i.e. neutral, and ending with an expression of maximal intensity. We will now define two metrics that will compare the emotion profiles of the neutral and peak expression images for each sequence in the CK+ database [22].

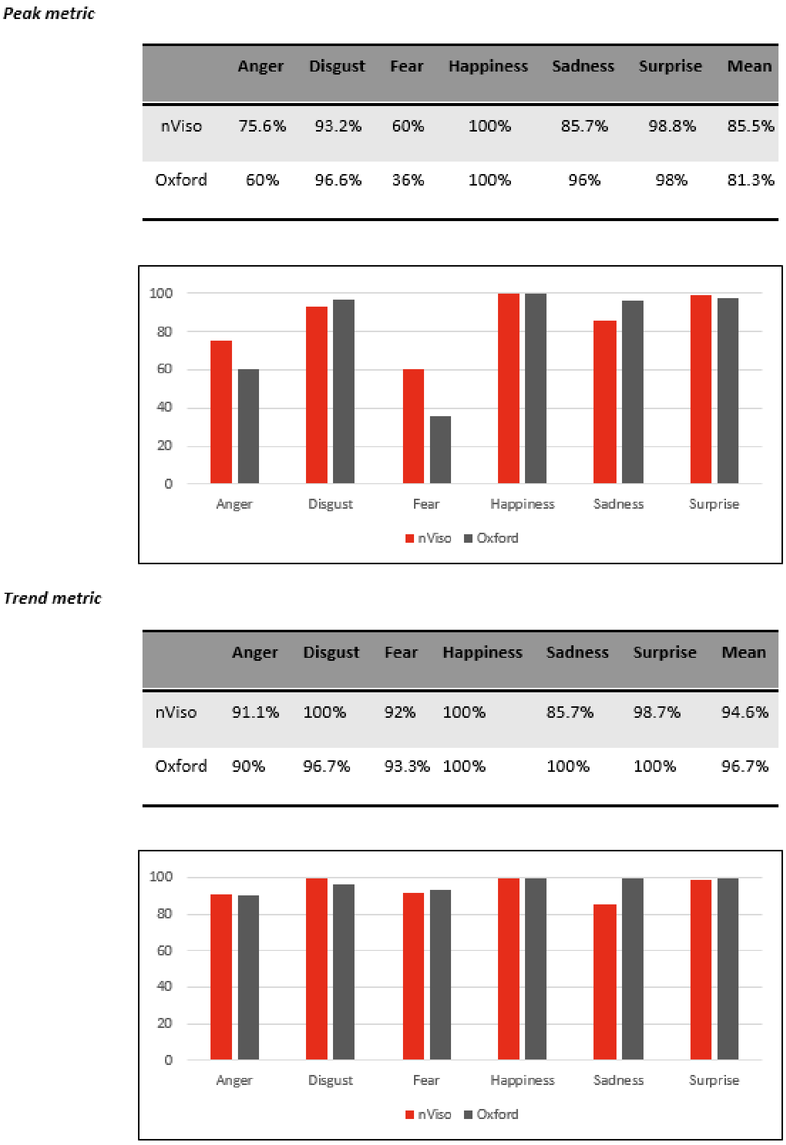
- Peak metric. To compute this metric, we compare the emotion profile of the maximum intensity expression with the emotion profile of the zero-intensity expression by creating a differential emotion profile which is simply the gain (or loss) of the level of activation of any emotion between the maximum intensity case and zero-intensity case. This metric is computed by considering the maximum value in the differential emotion profile and comparing the associated emotion to the

CK+ ground truth. If this maximum value corresponds to the ground truth emotion we consider the peak image correctly classified.

- Trend metric. To compute this metric, we consider the ground truth emotion of the peak image and we extract from its emotion profile the emotion intensity corresponding to that emotion. Same procedure is applied to the emotion profile for the neutral image. We then check the percentage gain and the absolute gain in these two emotion intensities and if the absolute gain is larger than 5% and the relative gain is larger than 50% we consider the peak image correctly classified.

3.2.4. Performance and Comparison

The results in Figure 7 correspond to the percentage of correctly classified images of the 593 sequences of the CK+ database considering the peak and trend metrics. In both cases the performance of the proposed model are in line with the state of the art approach.



Peak metric

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Mean
nViso	75.6%	93.2%	60%	100%	85.7%	98.8%	85.5%
Oxford	60%	96.6%	36%	100%	96%	98%	81.3%

Trend metric

	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Mean
nViso	91.1%	100%	92%	100%	85.7%	98.7%	94.6%
Oxford	90%	96.7%	93.3%	100%	100%	100%	96.7%

Figure 7. Comparison results for emotion inference in CK+ database using nViso and Oxford approaches.

4. Computational Performance of the Emotion Recognition Network

In this section, a comparative study between the two deep learning frameworks using the network for facial emotion recognition is described. The network was executed on a PC using tiny_dnn and in the EoT device with both the tiny_dnn and Fathom ports. The average inference time, average power consumption and energy needed per inference can be found in Table 1.

It is worth noting that the best results in terms of inference time were obtained using the Fathom framework regardless of the number of SHAVEs used. Although tiny_dnn is parallelised and gives computational times that make it suitable for real applications, the code has not been extensively optimised for the hardware architecture. On the contrary, Fathom was designed to be run in the Myriad 2 chip taking full advantage of its capabilities. The best absolute inference time was obtained using the Fathom framework with eight SHAVEs, 6.23 ms. Although it could be expected that an increase in the number of SHAVEs used would lead to a reduction in inference time, there are other factors like context changes or memory management that take time when more processors are involved.

Regarding power consumption, a key parameter in a portable battery-operated device, while the Fathom framework gives values ranging from 794.51 mW to 1387.71 mW, tiny_dnn requires 1100 mW. Due to the lower computational cost, it is important to carefully select the number of processors used considering a trade-off between inference time and energy accordingly to the application requirements.

Table 1. Emotion Recognition Network Performance.

Library	Platform	Time (ms)	Avg Power (mW)	Energy/Inference (mJ)
Tiny_dnn	PC (Core i7 4712HQ 2.30 GHZ)	190	N.A.	N.A.
Tiny_dnn	EoT (12 SHAVEs)	600	1100	660
Fathom	EoT (1 SHAVE)	14.6	794.51	11.6
Fathom	EoT (2 SHAVEs)	9.04	911.21	8.23
Fathom	EoT (3 SHAVEs)	8.48	971.75	8.24
Fathom	EoT (4 SHAVEs)	6.73	1067.74	7.19
Fathom	EoT (5 SHAVEs)	6.95	1105.68	7.69
Fathom	EoT (6 SHAVEs)	6.26	1172.11	7.34
Fathom	EoT (7 SHAVEs)	7.42	1149.14	8.52
Fathom	EoT (8 SHAVEs)	6.23	1254.21	7.82
Fathom	EoT (9 SHAVEs)	6.85	1242.69	8.51
Fathom	EoT (10 SHAVEs)	6.45	1308.41	8.45
Fathom	EoT (11 SHAVEs)	6.9	1307.46	9.02
Fathom	EoT (12 SHAVEs)	6.53	1387.71	9.061

5. Smart Doll

This smart doll illustrates the facial analysis system described above in the context of a real application (Figure 8). The EoT board was embedded inside a doll torso, performing facial emotion recognition so that the doll can assess the child's emotional display using deep learning and react accordingly with audio feedback (Figure 9). A face detector [23,24] is applied on each frame (it uses EoT's Inertial Measurement Unit (IMU) to rotate the input image so that a standard upright face detector can be applied), the face region is then cropped and resized to a 50 × 50 image which is then fed to the network described in the previous section. It is worth noting that all computations are local to the EoT device which reduces latency and power consumption and tackles privacy issues. The preserved privacy actually makes the application practical to the point of having commercial value. The emotional state of the infant can be recorded and this information can be also downloaded from the doll. As special care must be taken for the security of the recorded emotional state of the infant, the smart doll provides by default the possibility to store it in AES 128 bit encrypted form. This information can be only accessed when the doll is in configuration mode through the WiFi, using the password previously defined by the doll's owner. Two scenarios are considered in this application: as an interactive doll in a playful situation and as a therapeutic tool in the second.

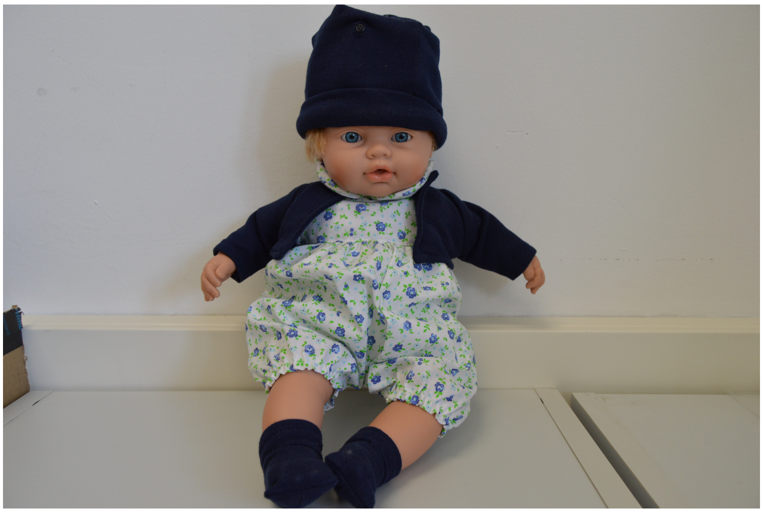


Figure 8. EoT doll.

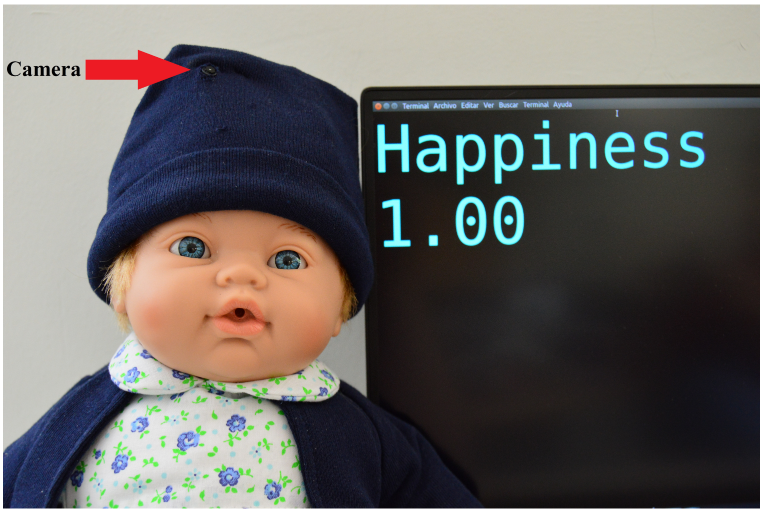


Figure 9. EoT doll detecting happiness emotion.

The doll incorporates speakers, a Sony IMX208 camera, which is currently located on the forehead but will be eventually placed on the eye of the doll in following versions to simulate the field of view of the doll and to improve the performance of the classification system, and a button to change the doll's operation mode. Each mode has a different LED representation (Figure 10). With a 3.7 V 4000 mAh battery of approximately 95 g of mass and a size of 37 × 19 × 68 mm (http://spanish.globalsources.com/gsol/I/Lithium-ion-battery/p/sm/1153316246.htm) the doll can becontinuously recognizing emotions for over 13 h with tiny_dnn and over 18 h with Fathom. This duration can be further extended using different techniques, such as auto-off modes, wake-up on IMU activity, etc. Moreover, the EoT board allows recharging the battery through a USB connector. A video of the doll recognizing emotions is available at https://flic.kr/p/Ws9dnm.

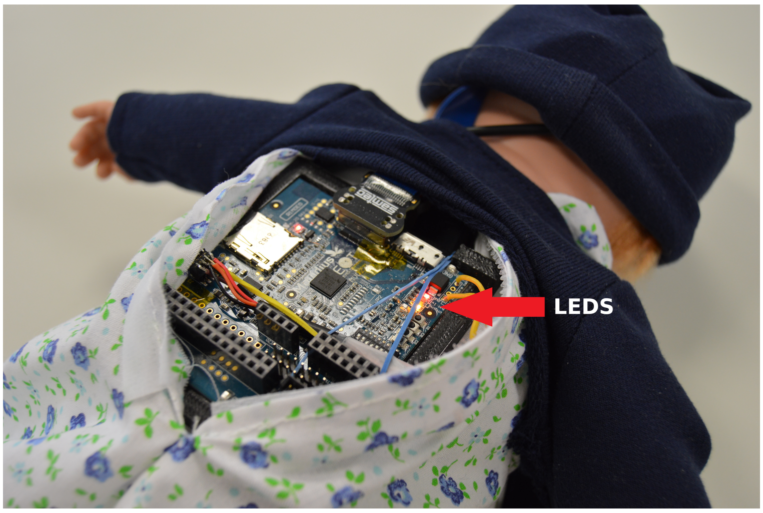


Figure 10. Inside the EoT doll.

5.1. Scenario 1—Interactive Doll–Talk to Me

The doll interacts with the infant through audio based on the perceived emotional state. Based on the detected emotion, the deep learning emotion engine will trigger predefined audio utterances. During play time, a LED on the EoT platform indicates that the doll is in its interactive analysis mode.

5.2. Scenario 2—Therapy Doll–I Can Help You

The doll is used to monitor and record the emotional behaviour of the infant while playing with the toy. In this scenario, the doll can be used in two different modalities:

- Recording mode: in this mode the doll will passively record the emotion while the infant plays with it.
- Playing-recording mode: in this mode the doll will record the session as in the previous mode and it will also provide audio feedback as in the first scenario. The recorded session will be downloadable through the Wi-Fi connection.

5.3. Smart Doll Main Application Structure

The standalone application developed for the doll flow graph is shown in Figure 11.

In the first steps the board is configured, loading both the face detector and emotion detector model files. After that, an initial welcome audio is played to inform the user that the doll is ready to be used.

From that point the demonstrator works in a loop capturing a frame in each iteration. The face detector is applied on each frame captured, and the face region detected is cropped and resized to a 50 × 50 image which is then fed to the network described in the previous section.

The doll considers that the user presents an emotion when it is recognised two consecutive times, and in this case, the application will print the result or play a feedback audio depending on the configuration of a Dual In-Line Package (DIP) switch of the EoT board.

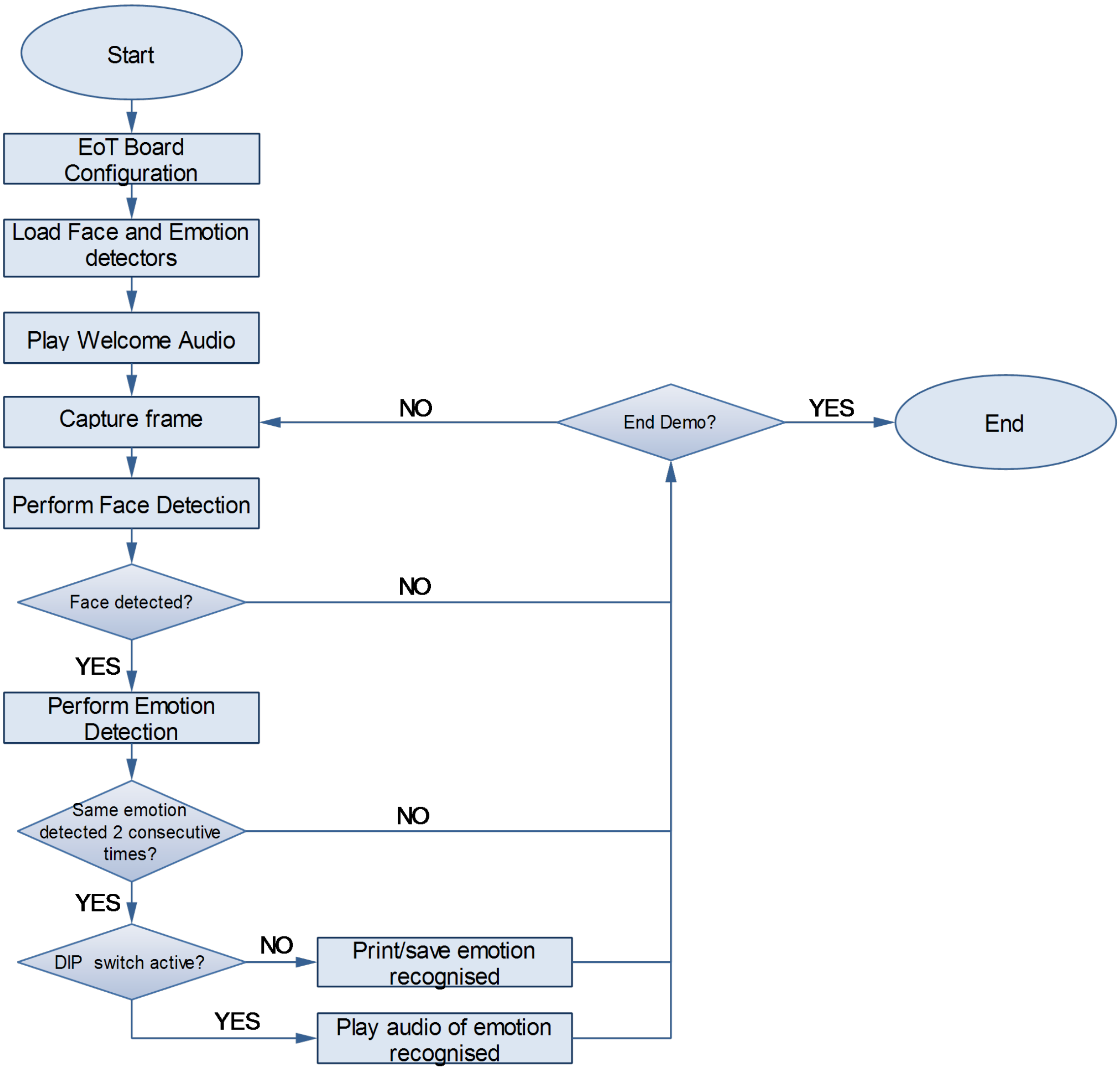

Start
EoT Board Configuration
Load Face and Emotion detectors
Play Welcome Audio
Capture frame
Perform Face Detection
Face detected?
YES
NO
Perform Emotion Detection
Same emotion detected 2 consecutive times?
YES
NO
DIP switch active?
NO
YES
Print/save emotion recognised
Play audio of emotion recognised
End Demo?
NO
YES
End

Figure 11. EoT doll flow diagram.

6. Pilot Test

The purpose of this pilot test was to explore the opportunities to use the EoT board empowered by an emotion recognition engine in a toy. The aim of the pilot was to test the system stability as mounted on the doll and to verify the accuracy of the emotion recognition application in correlation with the observations taken by researchers.

6.1. Location and Room Settings of the Pilot

The pilot was conducted on NVISO SA's office in Switzerland. A room was specifically prepared for the tests and equipped with: Smart Doll on a toy high chair; a children's chair facing the doll within 1 m distance; a camera behind the doll, facing the child once sitting on the chair; a chair for the facilitator and a desk for the researchers. This room is shown in Figure 12.

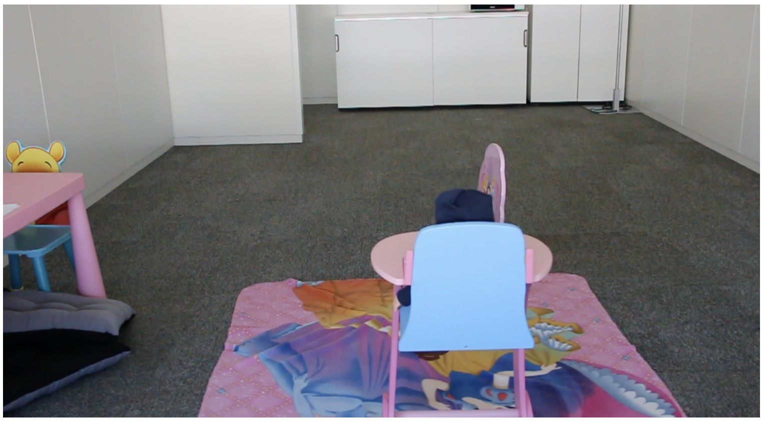
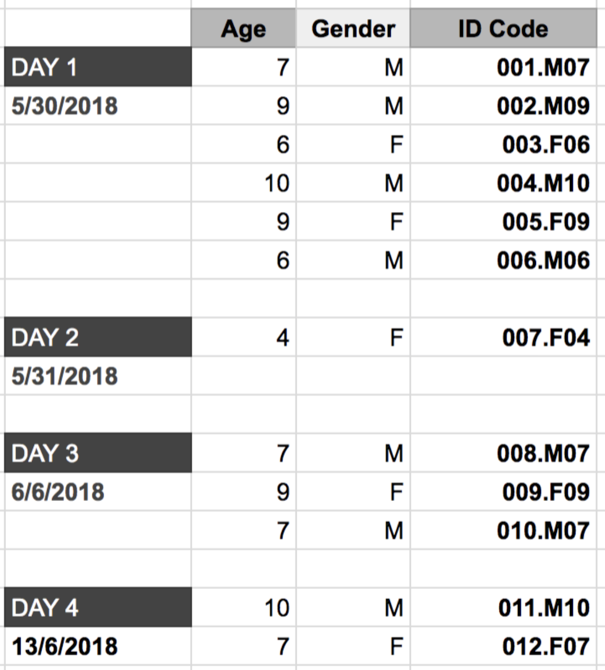


Figure 12. Room setup.

6.2. Participants

The tests were conducted on 12 children (Figure 13), aged between four and ten years old. The children are from various nationalities, and all of them are either French native speakers or are currently speaking French in everyday life (school). One child amongst the 12 children is affected by a mild form of autism. The only child aged four years old, once in the experience room, decided to not participate.



	Age	Gender	ID Code
DAY 1	7	M	**001.M07**
5/30/2018	9	M	**002.M09**
	6	F	**003.F06**
	10	M	**004.M10**
	9	F	**005.F09**
	6	M	**006.M06**
DAY 2	4	F	**007.F04**
5/31/2018			
DAY 3	7	M	**008.M07**
6/6/2018	9	F	**009.F09**
	7	M	**010.M07**
DAY 4	10	M	**011.M10**
13/6/2018	7	F	**012.F07**

Figure 13. List of children composing the panel (ID codes).

6.3. Pilot Description

Children were asked to sit in front of the talking doll for about 10 min. Each child was in the room with a coordinator and one or two researchers, and they were informed that they would be asked some questions. While the coordinator asked some generic question to the child, the doll greeted the child and introduced herself by surprise.

- Phase 1-Warm-Up. During the first warm-up phase, the doll asked three introductory questions (name, age, place of residence).

- Phase 2-Conversation. Once the child's attention was gathered, the doll asked a sequence of 12 questions. Each question was followed by a comment made by the doll. The content of the comment was triggered by the emotion detected by the doll's camera.
- Phase 3-Close. During the last phase of the experience, the doll asked the child if they had enjoyed interacting with her and bid farewell.

6.4. Results

The results are derived by comparing the emotion profiles recorded by the EoT board powered by the emotion engine with the annotations taken by the researchers during the test experiences.

The results in the section will focus on comparing the emotion profile to the researcher annotation, computing the accuracy and calculating the percentage of face detection. Missed faces are due to camera occlusion (e.g., the child has a hand on his face, or waves a toy in front of the camera) or absence of face (e.g., the child turns away from the camera).

The two examples used show different emotion behaviours: one with predominance of happiness, and the second with prevalence of neutrality, the latter is the emotion profile of the child affected by autism.

For the results we will use the legend of colours represented in Figure 14.

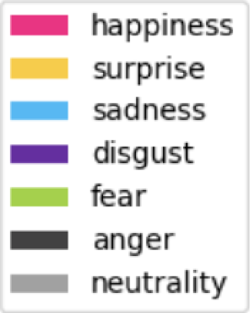


Figure 14. Legend of emotions. Best viewed in colour.

6.4.1. Video Processing Emotion Results

The images in Figures 15 and 16 show the emotion profile processed and recorded by the doll in two of the experiences conducted with two different children, ID 001 and ID 011. In the graphs, it is possible to see all of the seven emotions (including neutral) detected by the doll represented in different colours. Each vertical line corresponds to a point in time and the intensity of the emotion is proportional to the length of the coloured line.

As the graph shows, there are moments when more than one emotion is detected, all with different intensities. Figure 15 shows a predominance of happiness, while in Figure 16 neutrality is predominant. In Figure 16 missing data represents points in time where face was not detected due to child turning away his face from the doll's camera, or due to occlusions (objects between the child's face and the camera). The system had a percentage of successful face detection of 84.7%. Any time a face was detected, the image was processed and the resulting emotion was inferred.

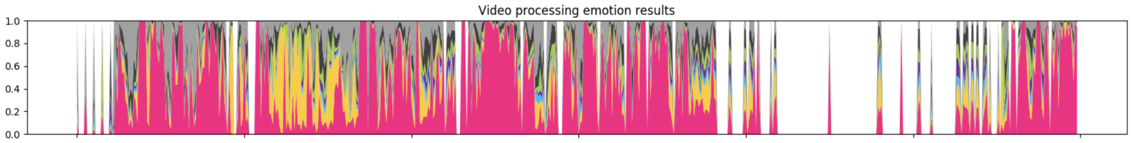


Figure 15. Emotion profile over time for child 001; *Y* = *Emotion intensity*, X = Time. Best viewed in colour.

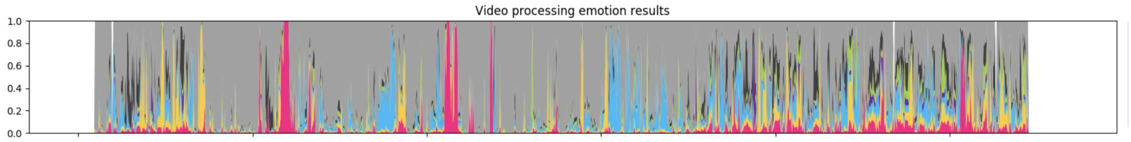


Figure 16. Emotion profile over time for child 011; *Y* = *Emotion intensity*, X = Time. Best viewed in colour.

6.4.2. Emotion Details

Figures 17 and 18 show the intensity curves of all emotions in seven separate graphs for children ID 001 and ID 011. A low pass filter is applied to smooth the lines. This representation allows to better identify and interpret the presence of multiple emotions at each point in time. For example, the presence of dominant happiness with a non-negligible level of surprise further highlights the engagement of the subject on the task he/she is performing, as in seconds 60 to 80 in the example in Figure 17. The same gaps in face detection shown in Figure 15 are visible also in the two graphs.

6.4.3. Comparison Between Doll Records and Researcher's Annotations

Figures 19 and 20 shows the comparison between the emotion profile processed and recorded by the doll's camera and the annotations taken by the researchers. Above the horizontal axis, it is possible to see the emotion profiles depicted, below the horizontal axis the graph of the researcher's annotations.

The vertical pink lines represent the points in time where an audio file was played by the doll. The first pink line corresponds to the first warm up question. As the two graphs show, different children interacted with the doll at different points, some children were more talkative, others gave more short and precise responses. This is true across the whole panel, with no significant impact on the average interaction time. The average interaction time from the first warm-up question asked by the doll to the last of the closing questions was about 4.41 min. The accuracy of the emotion detected compared to the note of the researchers for all the children was on average 83.5%.

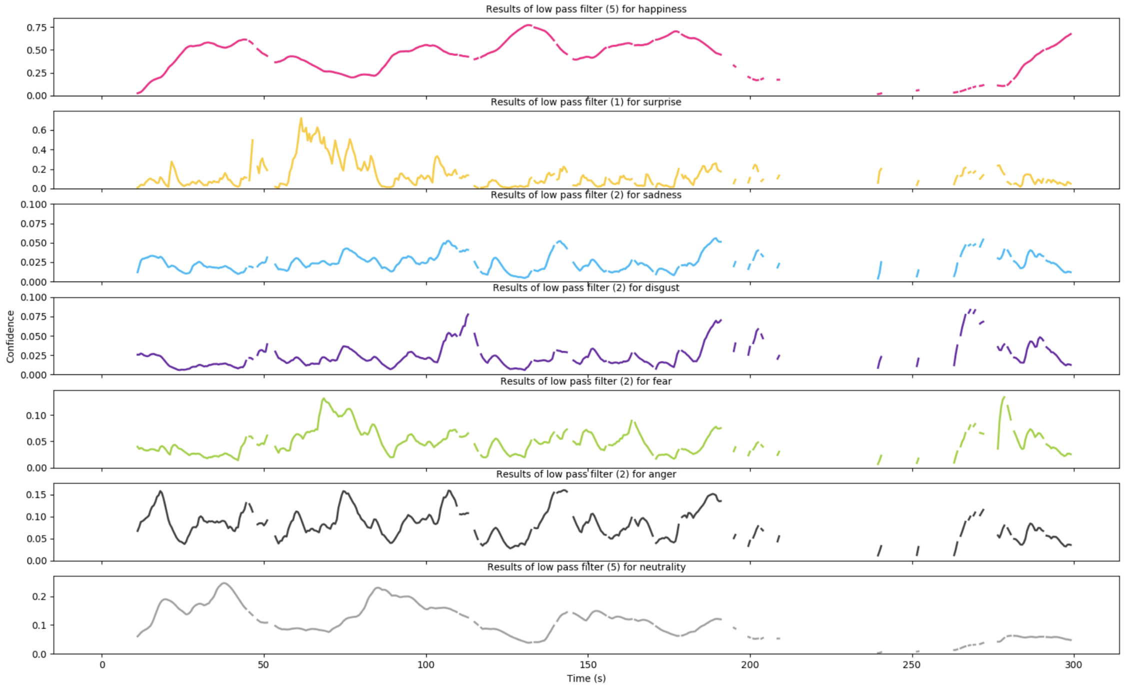


Figure 17. Emotion Curves over time for child 001. Best viewed in colour.

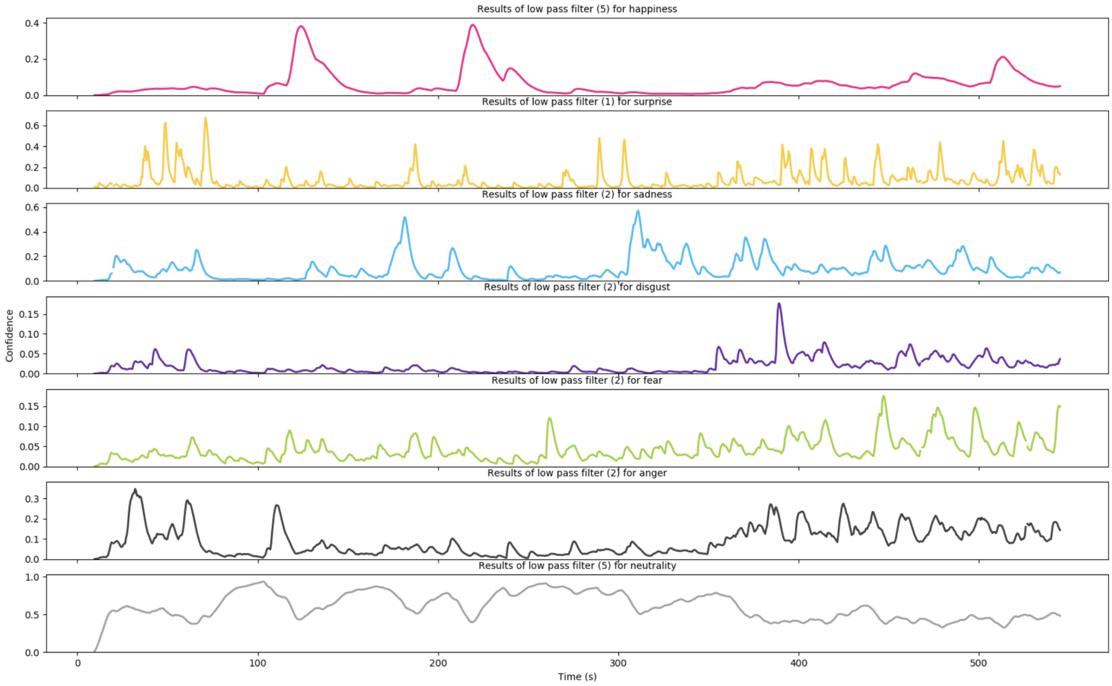


Figure 18. Emotion Curves over time for child 011. Best viewed in colour.

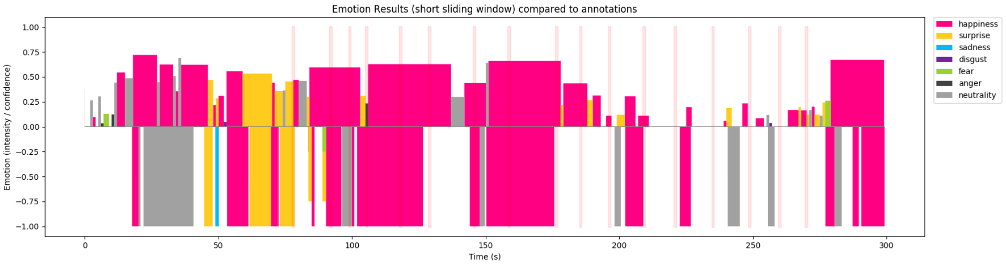


Figure 19. Emotion profile over time for child 001 compared to researcher's annotations. Best viewed in colour.

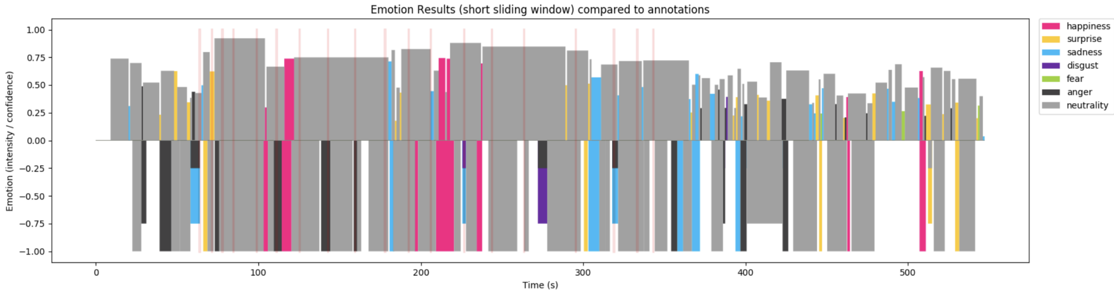


Figure 20. Emotion profile over time for child 011 compared to researcher's annotations. Best viewed in colour.

6.4.4. Children's Engagement

Figure 21 shows the percentage of the emotions detected across all the experiences. Well noticeable the predominance of happiness and neutrality. The first indicates engagement, while neutrality in association with peaks of happiness is correlated to punctual concentration. It is worth noting also that eight of the children claimed they would like to have a similar doll (or an interactively speaking toy) at home.

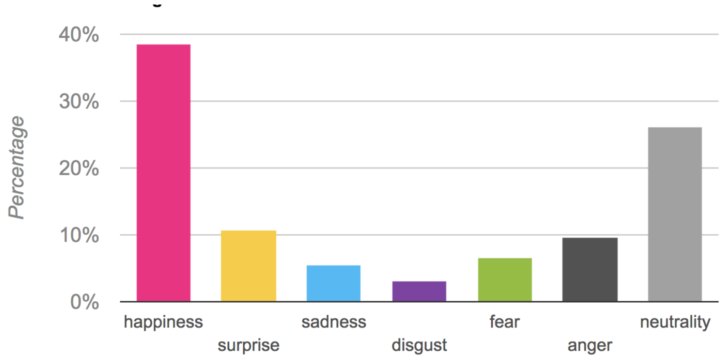


Figure 21. Average emotion predominance during the whole experience. Best viewed in colour.

6.5. Discussion

Based on these experimental results, we can state that the use of the proposed advanced application is viable in terms of real-time response using deep learning-based inference in an embedded system, and still the autonomy of the device makes it usable on a daily basis. Moreover, the local computation capability and the encryption methods provided by the device are able to effectively preserve the privacy of the owner and the infant, while the interest and the level of engagement of the children with the smart doll increased compared to a normal toy, allowing to explore its functionality as a therapeutic doll.

7. Conclusions

Eyes of Things is an open embedded computer vision platform. It was designed to optimize performance, power consumption, size, connectivity and cost. This makes it a flexible platform to implement vision-based applications. Apart from an extensive stack of computer vision and communications libraries, the potential of deep learning has been addressed in the form of CNN inference. We have used EoT in a real case of an emotional doll. Apart from performance, the privacy-preserving mode of operation, whereby no images are ever sent out of the toy, represent a crucial step forward.

A pilot was conducted to demonstrate the usability of the doll. The system was proved functional and stable and the average accuracy of the emotion engine was 83.5%. The pilot demonstrated a high level of engagement (emotion detected: surprise and happiness) of the children interacting with the smart doll, across all ages. The test on the software behaviour with the child affected by autism showed no difference in the level of accuracy of the emotion detected as compared to the researchers' observations. The results of the pilot suggest that the element of surprise generated by a doll that not only speaks, but also is able to interact and comment on what the children say, may be an attention catalyst across different children's ages. Overall our work is one of the first demonstrations that advanced vision capabilities can lead to novel engaging products.

Although most of the children claimed they would like to have a similar toy, one of the problems the smart doll developed will encounter is to convince parents about its security due to recent cases of data leaks in other toys on the market (https://www.forbes.com/sites/leemathews/2017/02/28/cloudpets-data-leak-is-a-privacy-nightmare-for-parents-and-kids/). Another limitation of the present study is the location of the camera on the forehead. As future work, an improved and more robust version of the doll will be developed placing the sensor in the eye, which could also improve the performance of the face detection and emotion recognition. Also, other low-power cameras will be tested to reduce the power consumption of the device and increase the battery duration. Moreover, an improved version of the proposed CNN will be developed to increase the ratio of emotions correctly recognised. Finally, further studies will be performed to assay the use of the doll as a therapeutic tool.

Author Contributions: Conceptualization, M.S.; Data curation, M.S.; Funding acquisition, G.B. and O.D.; Investigation, J.L.E.-A., N.V., G.B. and O.D.; Methodology, D.M. and O.D.; Project administration, O.D.; Resources, D.M.; Software, J.L.E.-A., N.V., J.M.R.-S., J.P.-P. and D.P.; Supervision, G.B. and O.D.; Validation, J.M.R.-S., J.P.-P. and D.P.; Writing–original draft, J.L.E.-A., N.V. and M.S.; Writing–review & editing, J.L.E.-A. and O.D.

Funding: This work has been supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 643924 [25].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Espinosa-Aranda, J.L.; Vallez, N.; Sanchez-Bueno, C.; Aguado-Araujo, D.; Bueno, G.; Deniz, O. Pulga, a tiny open-source MQTT broker for flexible and secure IoT deployments. In Proceedings of the 1st Workshop on Security and Privacy in the Cloud (SPC 2015), Florence, Italy, 30 September 2015; pp. 690–694.
2. Satyanarayanan, M.; Bahl, P.; Caceres, R.; Davies, N. The Case for VM-Based Cloudlets in Mobile Computing. *Pervasive Comput. IEEE* **2009**, *8*, 14–23. [CrossRef]
3. Sutaria, R.; Govindachari, R. Making sense of interoperability: Protocols and Standardization initiatives in IoT. In Proceedings of the 2nd International Workshop on Computing and Networking for Internet of Things, Mumbai, India, 8–9 November 2013.
4. Deniz, O.; Vallez, N.; Espinosa-Aranda, J.L.; Rico-Saavedra, J.M.; Parra-Patino, J.; Bueno, G.; Moloney, D.; Dehghani, A.; Dunne, A.; Pagani, A.; et al. Eyes of Things. *Sensors* **2017**, *17*, 1173. [CrossRef] [PubMed]
5. Intel® Movidius™ Myriad™ VPU 2: A Class-Defining Processor. Available online: https://www.movidius.com/myriad2 (accessed on 7 September 2018).
6. Barry, B.; Brick, C.; Connor, F.; Donohoe, D.; Moloney, D.; Richmond, R.; O'Riordan, M.; Toma, V. Always-on Vision Processing Unit for Mobile Applications. *IEEE Micro* **2015**, *35*, 56–66. [CrossRef]
7. Moloney, D.; Suarez, O. A Vision for the Future [Soapbox]. *Consum. Electron. Mag. IEEE* **2015**, *4*, 40–45, doi:10.1109/MCE.2015.2392956. [CrossRef]
8. Zhang, Y.; Wang, Y.; Zhou, G.; Jin, J.; Wang, B.; Wang, X.; Cichocki, A. Multi-kernel extreme learning machine for EEG classification in brain-computer interfaces. *Expert Syst. Appl.* **2018**, *96*, 302–310. [CrossRef]
9. Jiao, Y.; Zhang, Y.; Wang, Y.; Wang, B.; Jin, J.; Wang, X. A novel multilayer correlation maximization model for improving CCA-based frequency recognition in SSVEP brain–computer interface. *Int. J. Neural Syst.* **2018**, *28*, 1750039. [CrossRef] [PubMed]
10. Zhang, Y.; Zhou, G.; Jin, J.; Zhao, Q.; Wang, X.; Cichocki, A. Sparse Bayesian classification of EEG for brain–computer interface. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 2256–2267. [CrossRef] [PubMed]
11. Liu, N.; Wan, L.; Zhang, Y.; Zhou, T.; Huo, H.; Fang, T. Exploiting Convolutional Neural Networks with Deeply Local Description for Remote Sensing Image Classification. *IEEE Access* **2018**, *6*, 11215–11228. [CrossRef]
12. Wang, R.; Zhang, Y.; Zhang, L. An adaptive neural network approach for operator functional state prediction using psychophysiological data. *Integr. Comput. Aided Eng.* **2016**, *23*, 81–97. [CrossRef]
13. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117, doi:10.1016/j.neunet.2014.09.003. [CrossRef] [PubMed]
14. Tiny-Dnn. Avilable online: https://github.com/tiny-dnn/tiny-dnn (accessed on 7 September 2018).
15. Intel®Movidius™ Neural Compute Stick. Avilable online: https://developer.movidius.com/ (accessed on 7 September 2018).
16. Ekman, P.; Freisen, W.V.; Ancoli, S. Facial signs of emotional experience. *J. Pers. Soc. Psychol.* **1980**, *39*, 1125. [CrossRef]
17. Ekman, P. An argument for basic emotions. *Cognit. Emot.* **1992**, *6*, 169–200. [CrossRef]
18. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.

19. Microsoft Project Oxford Emotion API. Avilable online: https://www.projectoxford.ai/emotion (accessed on 7 September 2018).
20. Ekman, P.; Rosenberg, E.L. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*; Oxford University Press: Oxford, UK, 1997.
21. Hjortsjö, C.H. *Man's Face and Mimic Language*; Studentlitteratur: Lund, Sweden, 1969.
22. Goren, D.; Wilson, H.R. Quantifying facial expression recognition across viewing conditions. *Vis. Res.* **2006**, *46*, 1253–1262. [CrossRef] [PubMed]
23. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. I-511–I-518, doi:10.1109/CVPR.2001.990517.
24. Abramson, Y.; Steux, B.; Ghorayeb, H. Yet even faster (YEF) real-time object detection. *Int. J. Intell. Syst. Technol. Appl.* **2007**, *2*, 102–112. [CrossRef]
25. EoT Project. Avilable online: http://eyesofthings.eu (accessed on 7 September 2018).