Horizon 2020 PROGRAMME *ICT-01-2014: Smart Cyber-Physical Systems*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 643924

D4.10
Demonstrator 4
EoT Application

1. DOCUMENT INFORMATION

Authors	J.L. Espinosa-Aranda (UCLM) J.M. Rico (UCLM) N. Vallez (UCLM) J. Parra (UCLM) M. Sorci (nVISO) P. Sharma (nVISO) A. Pagani (DFKI)
Responsible Author	Dr. Alain Pagani (DFKI) e-mail: alain.pagani@dfki.de

Keywords	Demonstrator 4 – Empathic Doll
WP / Task	WP4
Nature	Other
Dissemination Level	PU

2. DOCUMENT HISTORY

Person	Date	Comment	Version
Alain Pagani	06.06.2018	Delivered version	1.0

3. ABSTRACT

In this deliverable, we describe the EoT application developed for the Empathic Doll Demonstrator.

This document presents the development process of the demonstrator and the requirements fulfilled and a brief explanation of the main application implemented. Moreover, the EoT libraries used are indicated.

4. TABLE OF CONTENTS

5. INTRODUCTION

This document describes the EoT Empathic Doll demonstrator main application and its use cases. In this demonstrator the application works in standalone mode without connecting it to any wireless network, and it does not allow to store the images captured, only the emotions detected in a text file, considering the ethical issues related to capturing child images. The main objective of the application is to detect the emotion of the person playing with the doll and provide some feedback/reaction through a speaker and save this information in the SD card for further study.

To carry out this objective the EoT device must contain the necessary files stored in the SD card, and can be configured while working using the dip switch 1 included on the EoT board.

The application can be found in the following repository:

https://gitlab.com/espiaran/EoT/tree/UCLM_FFBoard_mdk_17.04.5/WorkPackage_4/Smart_Doll/myriad/DollEoT

Additionally, another version of the demonstrator which includes an RTSP server has been developed only for demonstration purposes. It can be found in the following repository:

https://gitlab.com/espiaran/EoT/tree/UCLM_FFBoard_mdk_17.04.5/WorkPackage_4/Smart_Doll/myriad/DollEoT_rtsp

The reviewers will be able to access the private parts of the code on request.

6. SHORT DESCRIPTION OF THE DEMONSTRATOR

This demonstrator focusses on the implementation of an intelligent doll capable of analysing and interacting with an infant in an emphatic way.

Two scenarios have been considered in this demonstrator targeting a usage of the interactive doll in playful situation in the first scenario and as a therapeutic tool in the second.

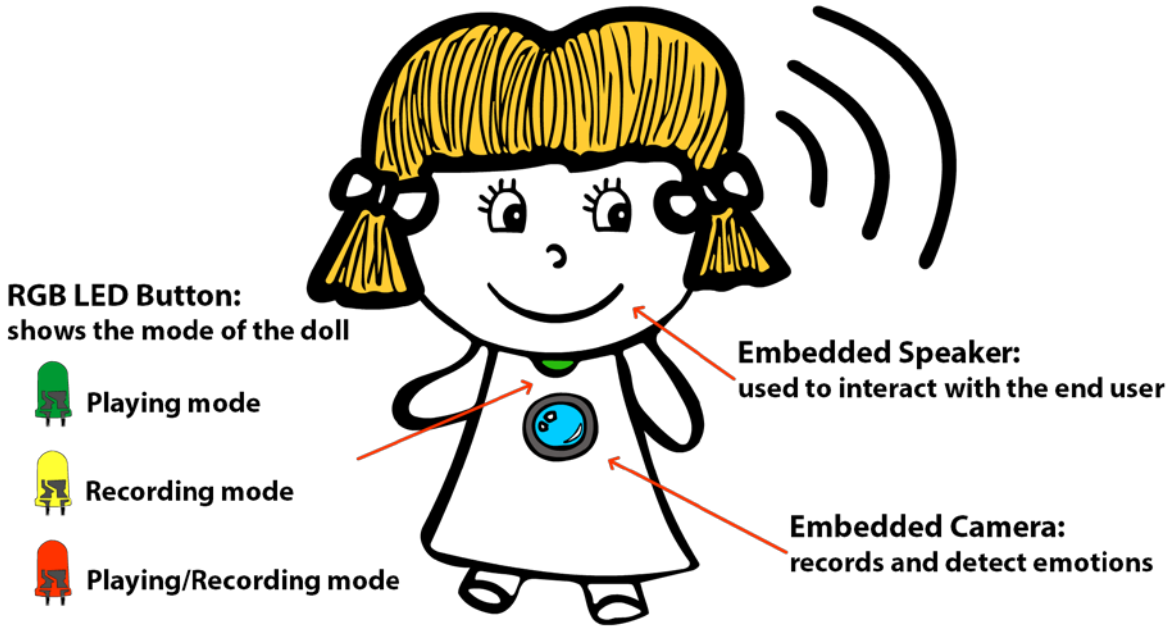


Figure 1: The Doll MockUp

The doll incorporates speakers, a camera and a backlight button to change the doll's mode. Every mode will have a different colour.

1. Scenario 1 – Interactive Doll – Talk to me

The doll will interact with the infant through an audio interactive feedback based on her emotional state. Based on her emotion the nViso emotion engine embedded in the EoT platform will trigger predefined audio comments. During the playing time, a green led on the EoT platform will indicate that the doll is in its interactive analysis mode.

2. Scenario 2 - Therapy Doll – I can help you

The doll will be used to monitor and record the emotional behaviour of the infant while playing with the toy. In this scenario, the doll can be used on 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 per previous mode and it will also provide audio interactive feedback as in the first scenario.

The recorded session will be downloadable through the Wi-Fi connection

7. EOT APPLICATION SOFTWARE DESCRIPTION

This demonstrator illustrates the deep learning capabilities implemented in EoT. 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. It is worth noting that all computations are local to the EoT device which reduces power consumption and tackles privacy issues. The preserved privacy makes the application practical to the point of having commercial value. The emotional state of the infant can be recorded if necessary encrypted and this information can be also downloaded from the doll.

Figure 2: The Doll

Figure 3: Embedded camera and emotion detected

The doll incorporates speakers, a camera and a button to change the doll's mode. Every mode has a different LED representation. The analysis and the recording of the face and of the emotion state can be performed for up to 13 hours in an always-on mode with a 3.7V 4000mAh battery. 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.

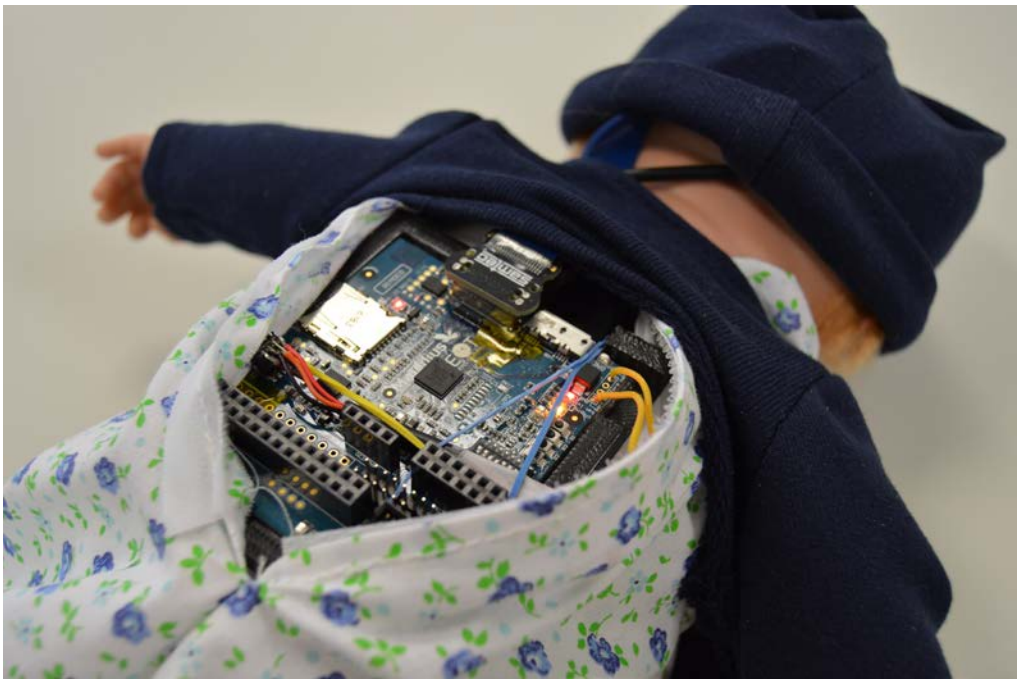


Figure 4: Back side of the Doll with EoT Board

1. Functional requirements

In this section, we present the requirements previously defined from the end-users' point of view and from a functional point of view. The following table recap the functional requirements and a brief description of the goal of each of them.

It is worth noting that these requirements have been fulfilled during the development of the demonstrator.

Table 1: List of functional requirements

ID	Name	Priority	Difficulty	Description
REQ001	Face detector	High	Medium	Cropped face of the detected infant. This is the output of the face detector processing the stream of images from the camera
REQ002	CNN module	High	High	Trained emotional module which gets the cropped face from the detector and provides the emotional profile as output
REQ003	Interactive Engine	High	Medium	The engine will get the inferred emotional profile and based on heuristics it will play an audio message
REQ004	Audio Feedback database	High	Medium	Database containing all possible audio interactions of the doll
REQ005	Saving sessions	High	Medium	Storing emotional profiles during entire sessions
REQ006	LED Switcher	Medium	Medium	Possibility to change the doll's mode (playing, recording, playing-recording) through a button which lights up in three different colours

REQ007	Session provider	Medium	Medium	Possibility to download the emotional profiles stored during sessions
REQ008	Module uploader	High	Medium	Upload the application into the doll

8. EOT APPLICATION SOFTWARE DOCUMENTATION

1. EoT libraries used

To develop the EoT device application of the Empathic Doll demonstrator, the following EoT libraries have been used:

- **SDCardIO**: This library is used to load the convolutional neural network from the SDCard
- **TimeFunction**: This library is used to update the time and date of the EoT device.
- **Camera**: This library is used to capture the images using the camera.
- **tiny_dnn**: tiny_dnn is a C++14 implementation of deep learning. It is suitable for deep learning on limited computational resource, embedded systems and IoT devices. This library is used to detect the facial expression of the captured images.
- **Audio**: This library is used play the audios stored on the SD Card depending on the emotion detected.
- **Libccv/OpenCV:** These libraries are used to work with the captured images. These libraries allow using computer vision algorithms, as face detection, image rotation, etc.
- **LEDs:** This library is used to manage the LEDs and the DIP switch used to control the behaviour of the doll.

2. Convolutional Neural Network(CNN) for emotion detection

1. Description of the emotion detection 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 analyzing the facial expressions, it is necessary to identify the basic set of emotions which are universal and can be quantified.

Dr. Ekman and his associates, studied people all around the world and even went to Papua New Guinea to study the “Fore” tribesmen, who have very limited contact with outsiders, to finally identify in 1971 the set of 6 universal basic emotions:

anger, disgust, fear, happiness, sadness and surprise.

Based on Dr. Ekman’s findings, humans are able to experience a very nuanced and wide panel of emotions, but few of them are said to be basic. To qualify an emotion as basic, it must respect eleven criteria, the 3 most important being:

1. Distinctive Universal Signals: The emotion needs to be facially expressible and signaling to observers that this specific emotion is experimented.

2. Distinctive physiology: The temporary physiological changes induced by the emotion should be distinguishable from other emotions.

3. Distinctive universals in antecedent events: Stimuli triggering the specific emotion must be distinguishable from stimuli associated with other emotions.

The nViso emotion engine is trained to classify the basic 6 emotions plus the neutral expression when no emotion is detected. Figure 5 provides an example of different emotion classes predicted by it.

Figure 5: Emotion classes predicted by nViso emotion engine

The nViso emotion engine uses a specially designed convolutional neural network(CNN) for emotion detection. In order to produces a model small enough to run on low powered embedded platforms like the one used for EoT, the network is trained on a relatively smaller dataset comprising of 6258 images obtained from 700 different subjects. The network is small with 7 RELU layers and 3 pooling layers.

Figure 6 depicts the architecture for CNN network used for training the emotion classifier.

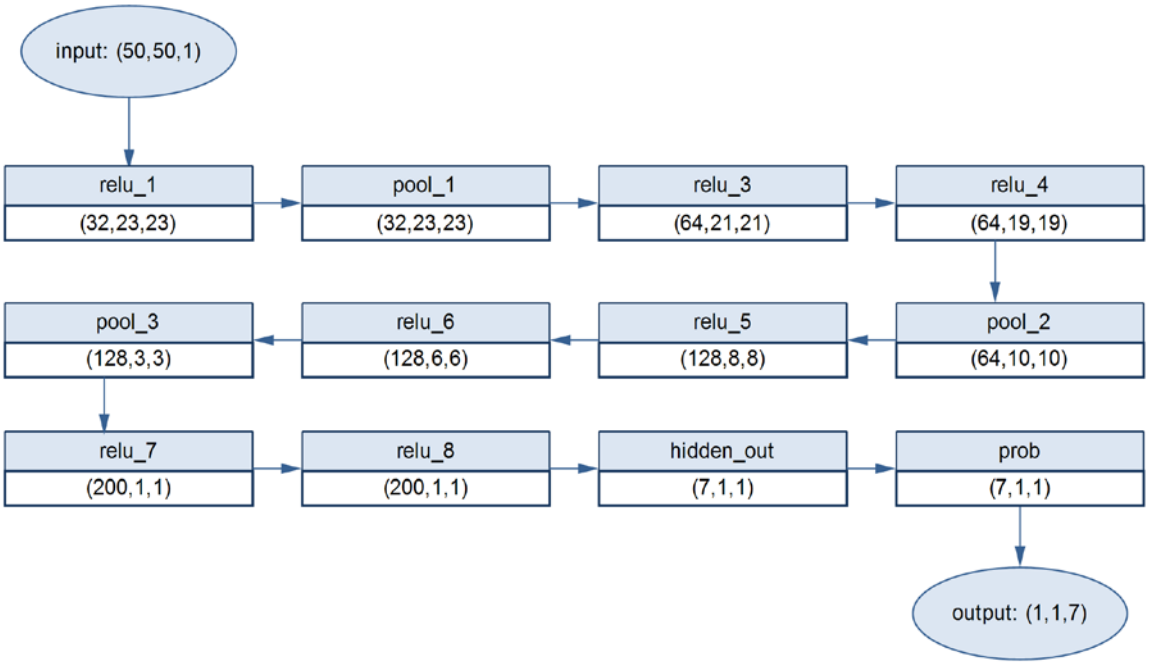


Figure 6: CNN architecture for emotion classification

As depicted in Figure 6, the input to network is an image of size 50x50 containing the face. The network applies a series of RELU and pooling layers with the output layer having 7 nodes one for each emotion. The trained network occupies a total of 900 KB of storage space and is stored on the EoT SD card.

2. Validation for the emotion model

For each validation procedure in machine learning a fundamental step is the choice of a testing database. The database should be chosen based on three fundamental requirements:

- Quality of the images.
- Quality of the annotations (ground truth).
- Usage of the database in the scientific community.

For the purpose of validation, we are considering one of the most widely used databases to benchmark the aforementioned performance – the so-called Cohn-Kanade extended database (CK+) which is described in detail in the next section. The next 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).

- **Benchmarking Database: Cohn Kanade Extended**

The Cohn-Kanade extended database (CK+) 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 proceed to a peak expression (an emotion), peak expressions are fully coded

with the Facial Action Coding System (FACS). Figure 7 provides an example of a sequence from CK+ dataset.

Figure 7: Example of a sequence from CK+ dataset

FACS is a system to taxonomize human facial expressions, the concept was originally invented by Swedish anatomist Hjortsjö in 1969, it was then developed and published in 1978 by Ekman and Friesen and improved once again in 2002 by Ekman, Friesen and Hager. FACS is currently the leading global standard for facial expressions 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) or one or several muscles, you can affine the coding by appending one of the 5 levels of intensities, which vary between “Trace” and “Maximum”, moreover every emotion is linked to a specific set of action units. Figure 8 provides some action annotations for a human face.

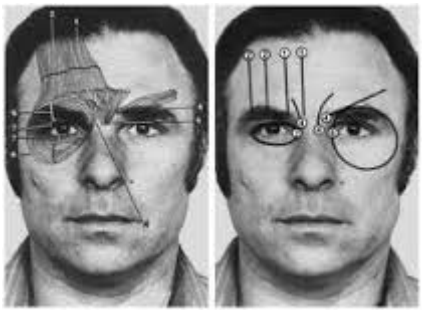


Figure 8: Action unit annotations for a human face

The fact that the peak expressions in the CK+ are FACS coded, facial expressions of very high quality, that many other databases do not have, ensuring the undeniable popularity of this dataset in the scientific research for the past 15 years.

- **Validation procedure**

Figure 9 represents the main steps from brain activation to output of the nViso engine. When a stimulus induces an emotion, the brain response causes the contraction of several facial muscles producing a facial expression which in turn represents macroscopically the activity inside our brain. Facial nerves establish a two-way communication between facial muscles and the activation of the amygdala and hippocampus. These two regions of the brain represent respectively the realm of the emotions and long-term memory. Through these facial muscle movements, the nViso emotion model is capable to accurately infer the emotional state of the person.

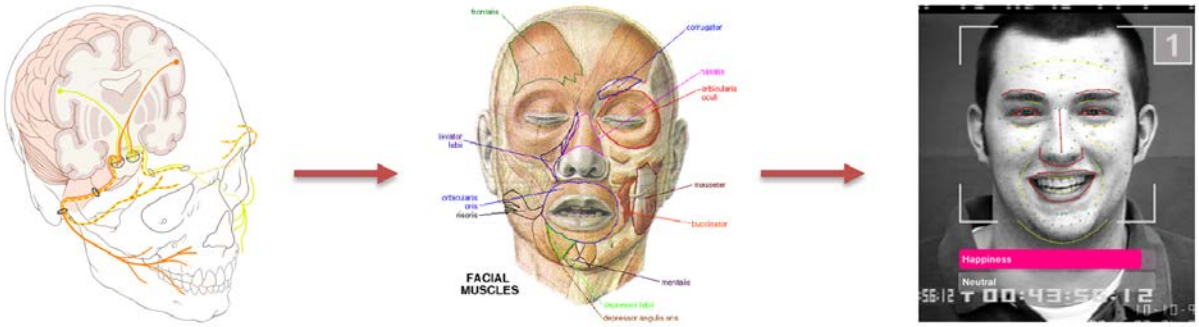


Figure 9: Brain to expressions to emotion inference

To analyse an image using the nViso engine, the positions of several key points on the face are detected using our face model. These key points are refined into a more insightful set of features and transformed by our emotion recognition engine into an emotion profile. 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, described in the next section, are computed.
To compare the performance with state-of-the-art approaches, the Oxford model introduced by Microsoft in 2015 is used. Oxford provides an online API to compute an emotion profile for a given image.

- **Validation Metrics**

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.

Peak metric

To compute the peak 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

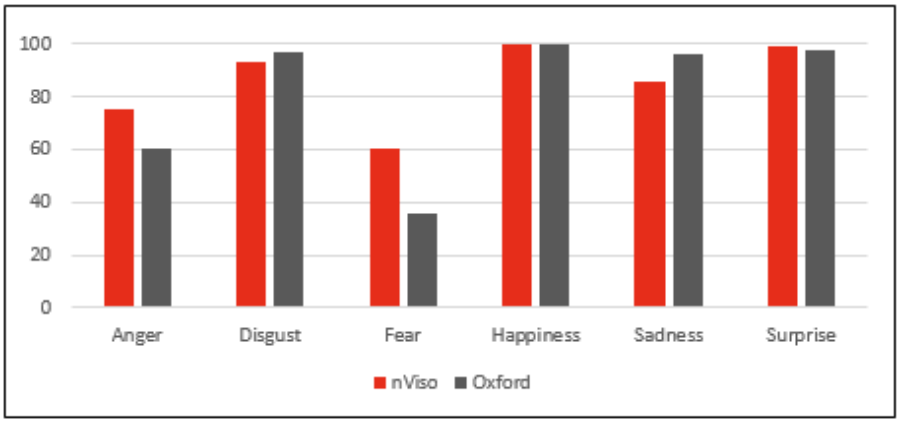
To compute the trend metric, we consider the ground truth emotion of the peak image and we extract from its emotion profile the emotion intensity corresponding to that emotion. Same procedure is applied to the emotion profile for the neutral image. We then check the percentage gain and the absolute gain 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.

- **Performance and comparison**

The results in Figure 10 correspond to the percentage of correctly classified image considering the peak and trend metrics.

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

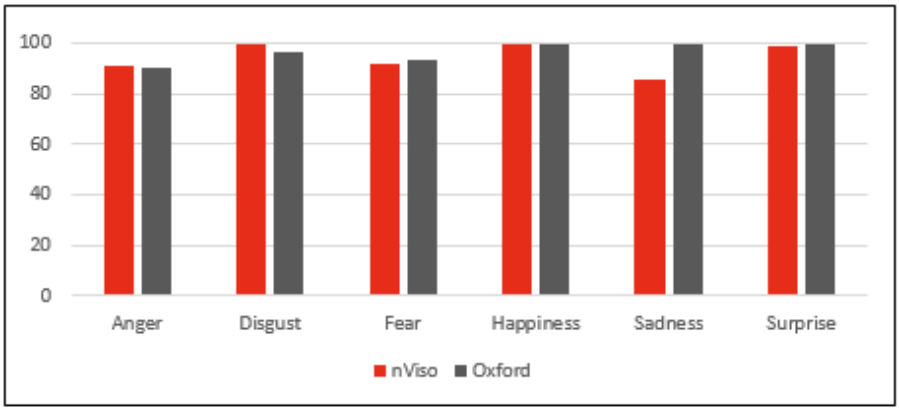


Figure 10: Comparison results for emotion inference

In both cases the nViso performance are in line with the state of the art approach.

3. EoT Doll main application structure

The standalone application developed for the doll flow graph is shown in the following picture.

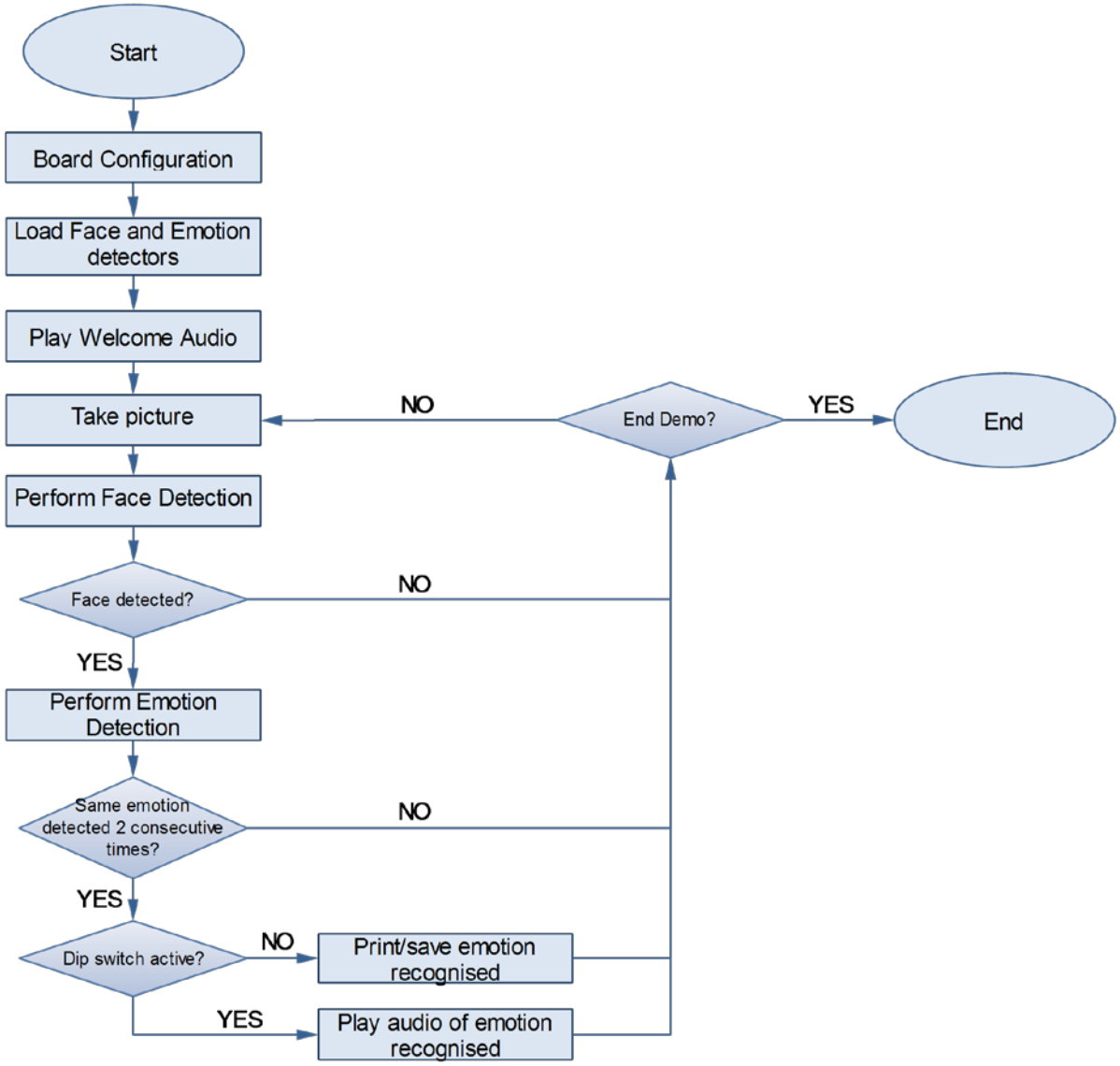


Figure 11: EoT doll demonstrator flow diagram

In the first steps the board is configured, loading both the face detector and emotion detector 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 taking a picture in each iteration. The face detector is applied on each frame captured (it uses EoT's IMU to rotate the input image so that an upright face detector can be applied), and the face region detected is cropped and resized to a 50x50 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 dip switch 1 of the EoT board.

9. CONCLUSIONS

This deliverable describes the EoT application used for the Empathic Doll Demonstrator where the privacy-preserving mode of operation, whereby no images are sent out of the toy, represent a crucial step forward.

The fulfilled requirements for the demonstrator has been described, as well as the libraries used, use cases and basic architecture of the application

- End of document -