Automatic Facial Expression Recognition for the Interaction of Individuals with Multiple Disabilities

Carmen Campomanes-Álvarez[1], Blanca R. Campomanes-Álvarez

[1]Ph.D. in Computer Science
Artificial Intelligence and Data Analysis Department
CTIC Technological Center - Gijón, Asturias, Spain
www.fundacionctic.org

ICAPAI
International Conference on Applied Artificial Intelligence

Outline

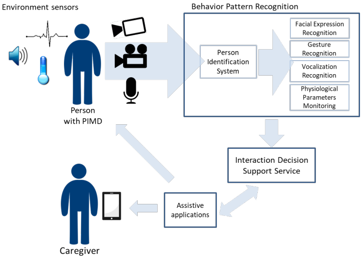
Introduction

People with Profound Intellectual and Multiple Disabilities (PIMD):

- present severe intellectual difficulties that affect their skill to communicate their feelings.
- can not be independent.
- express their needs with non-symbolic behaviors (non conventional reactions).
- need a constant support by a group of professional caregivers.

Introduction (II)

The INSENSION (H2020) project aims to design and develop an intelligent platform that enables people with PIMP to enhance the quality of their life with digital applications and services.



Based on the previous knowledge about each person, the platform will be able to associate the recognized expressions with their meaning.

Introduction (III)

Second stage in the INSENSION platform

Three stages:

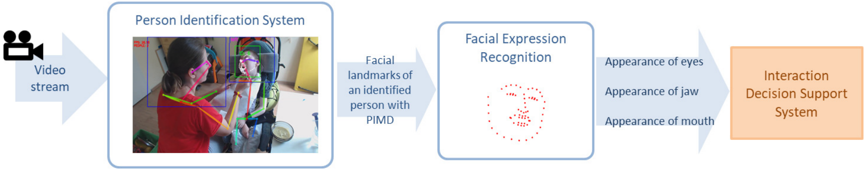
1. Person identification system
2. **Behaviour pattern recognition**
    - **Facial expression recognition**
    - Gesture recognition
    - Vocalization recognition
    - Physiological parameters monitoring
3. Interaction decision support service

Introduction (IV)

Facial Expression Recognition Component

The aim is to implement an automatic facial expression recognition system for:

- people with PIMD
- in real-time



Existing methods for facial expression recognition

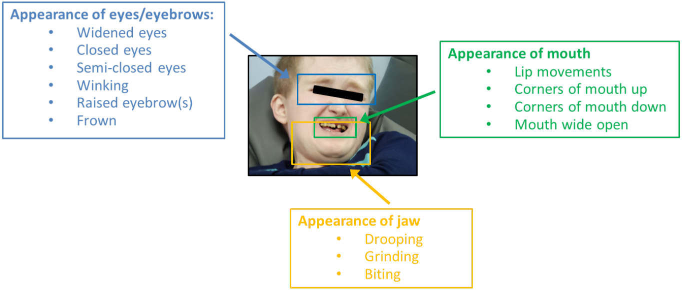
Automatic facial recognition approaches

- Automatic Facial Expression Analysis refers to systems that automatically analyze and recognize facial motions and facial feature changes from images.
- Current trends aim to recognize facial actions and their corresponding emotions using ML and CV techniques (basic emotions).
- The output usually is a label of an emotion or facial action with their intensity.

Our proposal: recognition per region

People with PIMD are not able to communicate or express their feelings with the well-known expressions.

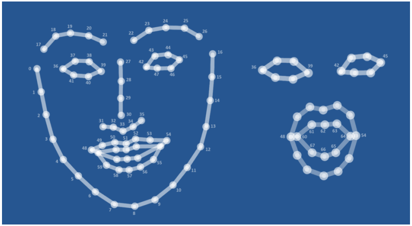
For the INSENSION project, it is necessary to analyse the facial expressions by region which were selected by pedagogical experts



Methodology

Common methodology

- 2D-to-3D Facial Landmarks Networks Baseline
  - estimation of the 3D coordinates from the 68 facial points detected in a 2D image thanks to a deep neural net.
  - location of the coordinates x, y, and z.
  - the score of the detection of each facial landmark indicates the confidence of the estimation.



v: video stream provided by one camera (RGB data) at time t:

$$v(t) \rightarrow I(t) \forall t = 1, ..., T, \qquad (1)$$

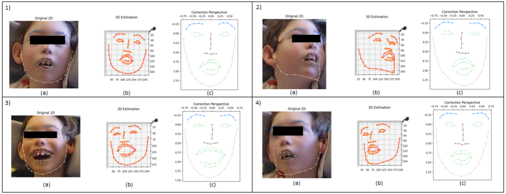
$$I(t) = \{(x(t), \; y(t), \; z(t), \; s(t))\}_{j \in J}, \qquad (2)$$

$$J = \{0,..,67\}$$

Methodology (II)

Common methodology

- Perspective Correction



- Facial landmarks centring

$$(\overline{x}_j, \overline{y}_j) = (x'_j, y'_j) - \left(x'_{27}, y'_{27}\right) \;\; \forall\, \mathrm{j} \in J. \tag{3}$$

- Facial landmarks scaling

$$\overline{\overline{x}}_j = \frac{\overline{x}_j}{\|\overline{v}_{36,45}\|}, \overline{\overline{y}}_j = \frac{\overline{y}_j}{\|\overline{v}_{36,45}\|} \tag{4}$$

Methodology (III)

Common methodology - Output

- Value between 0 and 1, which indicates the degree of confidence of doing each facial expression.

$$\mu_f\left(t\right), \tag{5}$$

- Visibility of facial regions ranges from 0 to 1

$$w_f\left(t\right) = \frac{1}{n}\sum_{j=k}^{m} s(t)_j, \tag{6}$$

where n is the length of L_f, k corresponds to the first element of the L_f set, and m to the last one.

Methodology (IV)

Classification Methods considered

Three types of approaches depending on the complexity of the task

- Logic reasoning of the anatomical performance (eyes appearance)
- ML classification models (eyebrows and mouth appearance)
  - Gaussian Naïve Bayes
  - K-Nearest Neighbor (KNN)
  - Logistic Regression
  - Stochastic Gradient Descent (SGD)
  - Neural Networks
  - Random Forest
- Long-Short Term Memory (LSTM) Neural Nets for sequence learning (jaw movements)

Methodology (V)

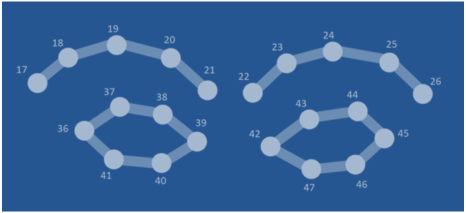
Evaluation methodology

- weighted accuracy
- weighted average precision
- weighted average recall
- weighted average F1-score

Methodology (VI)

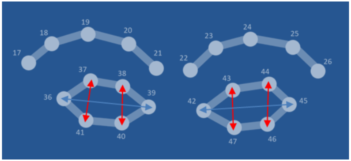
Appearance of the Eyes

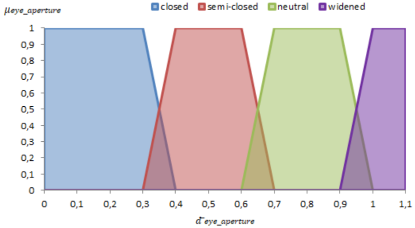
- Expressions:
  - widened eyes
  - semi-closed eyes
  - closed eyes
  - winking
  - *neutral - open*



Methodology (VII)

Appearance of the Eyes (I)





$$d_{eyeap_{right}}(t) = \max\left\{\left\|I(t)_{37} - I(t)_{41}\right\|,\ \left\|I(t)_{38} - I(t)_{40}\right\|\right\}, \quad (7)$$

$$d_{eyeap_{left}}(t) = \max\left\{\left\|I(t)_{43} - I(t)_{47}\right\|,\ \left\|I(t)_{44} - I(t)_{46}\right\|\right\}, \quad (8)$$

$$d_{ref_{right}}(t) = \frac{\left\|I(t)_{36} - I(t)_{39}\right\|}{2}, \quad (9)$$

$$d_{ref_{left}}(t) = \frac{\left\|I(t)_{42} - I(t)_{45}\right\|}{2}. \quad (10)$$

$$\overline{d}_{eye\_aperture} = \frac{d_{eye\_aperture}}{d_{ref}} \quad (11)$$

Methodology (VIII)

Appearance of the Eyes (II)

- Aggregation of both eyes

$$\mu_{ap}(t) = \frac{\mu_{ap_r}(t) \cdot \omega_{eye_r}(t) + \mu_{ap_l}(t) \cdot \omega_{eye_l}(t)}{\omega_{eye_r} + \omega_{eye_l}} \tag{12}$$

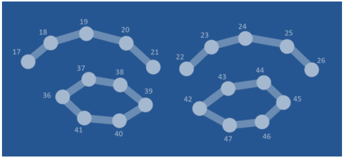
The action "winking" is performed when an eye is open (neutral) meanwhile the other one is closed.

$$\mu_{winking}(t) = \max\{\mu_{neutral_r}(t) \cdot \mu_{closed_l}(t), \mu_{neutral_l}(t) \cdot \mu_{closed_r}(t)\}. \tag{13}$$

Methodology (IX)

Appearance of the Eyebrows

- Expressions:
  - raised
  - frown
  - *neutral*



- The input vector for training the ML classifiers is composed by the positions of the five landmarks of each eyebrow

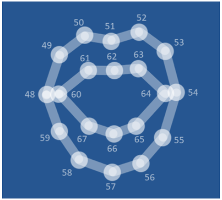
$$v_{eyebrow_{r|l}} = \left(l_j\left(t\right)\right),\ j\ \in\ L_{eyebrow_{r|l}}, \tag{14}$$

being $L_{eyebrow_r} = \{17,\ \ldots,\ 21\}$ and $L_{eyebrow_l} = \{26,\ \ldots,\ 22\}$

Methodology (X)

Appearance of the Mouth

- Expressions:
  - lips movements
  - corners of mouth up
  - corners of mouth down
  - mouth wide open
  - *neutral*



- The input vector for training the ML classifiers is composed by the positions of the twelve landmarks of each eyebrow

$$v_{mouth} = (l_j(t)), \ j \ \in \ L_{mouth}, \qquad (15)$$

being $L_{mouth} = \{48, \ \ldots, \ 60\}$

Methodology (XI)

Appearance of the Jaw

- Expressions:
  - drooping
  - grinding
  - biting
  - *neutral*
- The input vector for training the LSTM NN classifier is composed by some distances between chin landmarks for a sequences of 20 samples:

$$
\begin{aligned}
v_{jaw} = \{ & d\left(l_{62}(t),\ l_{66}(t)\right), \left(x_{51}(t) - x_{57}(t)\right), \\
& \left(y_{51}(t) - y_{57}(t)\right), \left(x_{51}(t) - x_{59}(t)\right), \\
& \left(y_{51}(t) - y_{59}(t)\right), \left(x_{51}(t) - x_{55}(t)\right), \\
& \left(y_{51}(t) - y_{55}(t)\right)\}\ t \in \{0, \ldots 20\}.
\end{aligned} \tag{16}
$$

Experiments and results

General setup

- Intel CoreTM i7-5820K CPU 3.30GHz x 12 with 31.3 GiB RAM, TITAN Xp Graphic Card, Ubuntu 16.04LTS.
- Facial landmarks are detected in 2D and estimated in 3D by Face Alignment API.
- Python libraries scikit-learn and Keras for training and test the classification methods.

Experiments and results (II)

Datasets

Eyes appearance	Total samples
Closed	75
Semi-closed	75
Widened	75
Winking	75
Neutral	75

Frown appearance	Total samples
Frown	75
Raising	75
No-movement	75

Mouth appearance	Total samples
Corners of mouth up	74
Corners of mouth down	84
Mouth wide open	70
Lips movements	65
Neutral	82

Jaw movement	Total samples
Grinding	339
Biting	370
Drooping	370
No-movement	299

Experiments and results (III)

Results for Eyes Appearance

	Predicted appearance				
Real appearance	**Closed**	**Semi**	**Widened**	**Winking**	**Neutral**
Closed	0.850	0.110	0.000	0.040	0.000
Semi	0.040	0.880	0.000	0.005	0.030
Widened	0.000	0.000	0.850	0.000	0.150
Winking	0.000	0.050	0.030	0.910	0.010
Neutral	0.010	0.040	0.040	0.010	0.900

Accuracy	**Precision**	**Recall**	**F1-score**
0.89	0.89	0.89	0.89

Experiments and results (IV)

Results for Eyebrows Appearance

Method	Accuracy	Precision	Recall	F1-score
Neural Network	0.90	0.90	0.90	0.90
Random Forest 150	0.80	0.81	0.80	0.80
K-Nearest Neighbour	0.76	0.77	0.76	0.77
SGD	0.76	0.77	0.76	0.74
Random Forest 100	0.76	0.76	0.76	0.76
Random Forest 50	0.75	0.74	0.75	0.74
Naïve Bayes	0.73	0.72	0.73	0.72
Logistic Regression	0.71	0.71	0.71	0.70

	Predicted appearance		
Real appearance	**Frown**	**Raising**	**Neutral**
Frown	0.85	0.00	0.15
Raising	0.00	1.00	0.00
Neutral	0.13	0.00	0.87

Experiments and results (V)

Results for Mouth Appearance

Method	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbor	0.90	0.90	0.90	0.90
Random Forest 150	0.88	0.91	0.88	0.88
Random Forest 100	0.88	0.90	0.88	0.88
Random Forest 50	0.88	0.88	0.88	0.88
Naïve Bayes	0.82	0.86	0.82	0.82
Logistic Regression	0.86	0.89	0.86	0.85
SGD	0.86	0.86	0.86	0.86
Neural Network	0.72	0.72	0.72	0.71

	Predicted appearance				
Real appearance	**Corners up**	**Corners down**	**Wide**	**Lip mov.**	**Neutral**
Corners up	0.93	0.00	0.00	0.00	0.07
Corners down	0.00	0.89	0.00	0.00	0.11
Wide	0.00	0.00	1.00	0.00	0.00
Lips mov.	0.00	0.00	0.00	0.67	0.33
Neutral	0.00	0.17	0.00	0.00	0.83

Experiments and results (VI)

Results for Jaw Appearance

nº IL, nº of HN	Acc.	Precision	Recall	F1
2 layers: 280-140	0.80	0.82	0.80	0.80
3 layers: 280-560-140	0.73	0.74	0.73	0.72
4 layers: 280-560-280-140	0.72	0.74	0.72	0.73
5 layers: 280-560-280-140-70	0.68	0.71	0.68	0.69
6 layers: 280-560-840-560-280-140	0.72	0.73	0.72	0.71

	Predicted appearance			
Real appearance	**Grinding**	**Biting**	**Drooping**	**Neutral**
Grinding	0.74	0.06	0.03	0.17
Biting	0.05	0.80	0.00	0.16
Drooping	0.25	0.00	0.75	0.00
Neutral	0.00	0.04	0.00	0.96

Conclusions

- This component implies the development of a new approach, which analyzes these changes by region, unlike the current methods, which are based on identifying known expressions of the whole face.
- The general approach follows some strategies for unifying the face to assure the robustness to camera distance, subject movements, different body sizes and perspective.
- The proposed system is composed by four subsystems, i.e., eyes, eyebrows, mouth and jaw, obtaining high-quality results for the required application.

ICAPAI

International Conference on Applied Artificial Intelligence

Thank you for your attention!

Carmen Campomanes Álvarez, Ph.D.
CTIC Technology Centre, Gijón, Asturias, Spain.

carmen.campomanes@ctic.es

CTIC centro tecnológico