Recognition of Behaviour Patterns for People with Profound Intellectual and Multiple Disabilities

Erik Dovgan
Department of Intelligent Systems
Jožef Stefan Institute
SI-1000 Ljubljana, Slovenia
erik.dovgan@ijs.si

Jakob Valič
Department of Intelligent Systems
Jožef Stefan Institute
SI-1000 Ljubljana, Slovenia
jakob.valic@ijs.si

Gašper Slapničar
Department of Intelligent Systems
Jožef Stefan Institute
SI-1000 Ljubljana, Slovenia
gasper.slapnicar@ijs.si

Mitja Luštrek
Department of Intelligent Systems
Jožef Stefan Institute
SI-1000 Ljubljana, Slovenia
mitja.lustrek@ijs.si

ABSTRACT

People with profound intellectual and multiple disabilities (PIMD) are hard to understand because they are not capable of symbolic communication. Artificial intelligence can play a key role in recognizing behavior patterns with which they express themselves. It can thus assist new caregivers that are not familiar with a PIMD person. Within the INSENSION project, we developed a behavior pattern recognition approach that classifies a person's inner states and communication attempts based on his/her facial expressions, gestures, vocalizations, and physiological signals.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Artificial intelligence**.

KEYWORDS

Decision Support System; Pattern Recognition; People with PIMD

ACM Reference Format:
Erik Dovgan, Jakob Valič, Gašper Slapničar, and Mitja Luštrek. 2021. Recognition of Behaviour Patterns for People with Profound Intellectual and Multiple Disabilities. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp-ISWC '21 Adjunct), September 21–26, 2021, Virtual, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3460418.3479370

1 INTRODUCTION

People with profound intellectual and multiple disabilities (PIMD) have cognitive and physical disabilities as well as great difficultly communicating. They typically do not use symbolic communication, but communicate with facial expressions, vocalizations and body language that is different for each person with PIMD. As a consequence, their communication is hard to understand and even skilled caregivers need time to learn their communication patterns.

Several artificial intelligence methods have been developed for recognizing a person's body movements, facial expressions, gestures, etc. from video. However, basic features such as facial expressions and body movements are not enough since complex relations between them might exist. Additional challenges are that these features may have a different meaning for each person and are not used consistently.

To discover complex communication relations between basic features we developed a decision support system (DSS) that recognizes behavior patterns related to person-specific inner states and communication attempts. DSS processes facial expressions, gestures, vocalizations, and physiological data to classify inner states and communication attempts, and combines the obtained classifications with context data such as objects and other persons in the room in order to obtain contextualized decisions. The developed approach is part of the INSENSION project [2].

We present the implementation of the DSS for two persons with PIMD. The sensing system was installed in the kindergarten they attend and used during their regular activities with caregivers (kindergarten personnel). The persons with PIMD were not aware of our experiment, and the caregivers were focusing on their task, so the experiment closely resembled real-life usage of the system.

2 BEHAVIOR PATTERN RECOGNITION APPROACH

The behavior pattern recognition approach consists of preprocessing the input data into recognized gestures, facial expressions, vocalizations, and physiological states, building decision models that classify inner states and communication attempts, combining the decisions with context data, and updating decision models based on information from caregivers. The model-updating procedure is based on the active learning approach and is not described in this paper, while details on INSENSION's facial expression recognizer, gesture recognizer, physiological state recognizer, and vocalization recognizer can be found in [3, 5].

UbiComp-ISWC '21 Adjunct, September 21–26, 2021, Virtual, USA

ACM ISBN 978-1-4503-8461-2/21/09…$15.00
https://doi.org/10.1145/3460418.3479370

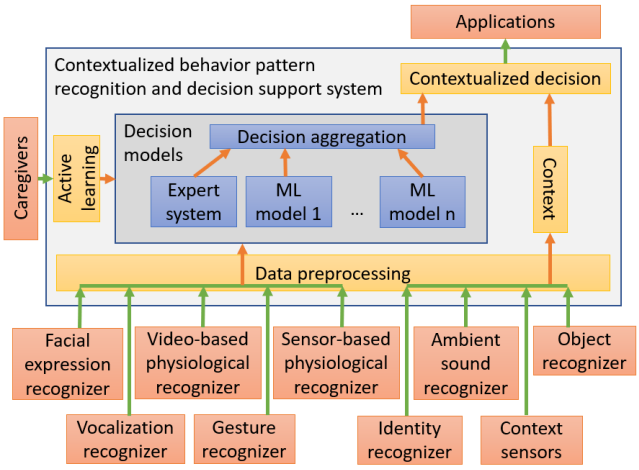

Figure 1: Contextualized behavior pattern recognition and decision support system architecture.

The architecture of the decision support system is shown in Figure 1. At the bottom of the figure are two sets of inputs. The first set is related to the target person only and uses sensor readings (from cameras, microphones and the Empatica E4 physiological monitoring wristband) to extract the person's non-symbolic behaviour signals. The second set uses sensor readings (IoT system, cameras and microphones) to extract important information about the person's context. Additional context information about other people in the vicinity is derived from the identity recognizer.

The input data are first preprocessed by collecting messages from the recognizers related to a time interval of interest (set to 10 s) and extracting features characterizing the target person's behavior and context. These features are assembled into an instance, the basic unit of information for decision making. Such instances are fed into decision models, which output the person's inner state (pleasure, displeasure and neutral) and communication attempt (comment, demand, protest and none). Since the decision models can never be made perfect, and people can change with time, we incorporated active learning into the system. This means that in case a caregiver notices a wrong decision by the system, he/she can correct it, so the decision models are retrained using the new information. The context arriving in the DSS is merged with the recognized inner states and communication attempts into contextualized decisions. These are finally used by assistive applications to take appropriate actions (e.g., communicate the target person's inner state to the caregivers, or turn media player on/off).

2.1 Dataset and Data Preprocessing

Within the INSENSION project, data from two PIMD persons were collected and annotated with inner state and communication attempt. The study was approved by the Bioethical Committee at the Poznan University of Medical Sciences (approval number 10/21) following international standards.

Basic statistics of the obtained 10-second instances are shown in Tables 1–2. These data are quite imbalanced with respect to the classes. This is mostly because the objective of caregivers is to prevent displeasure, protest and to some degree demand, and it would not be ethical to elicit these inner states and communication attempts on purpose. However, imbalanced data is a problem for machine learning, because machine-learning algorithms build models that favor better represented classes. We tackled this problem with two class balancing methods.

Table 1: Instances Labelled with Inner States

Inner state	All instances		Instances with physiological data	
	Person 1	Person 2	Person 1	Person 2
Neutral	561	307	291	307
Pleasure	102	154	18	154
Displeasure	82		27	

Table 2: Instances Labelled with Communications Attempts

Communication attempt	All instances		Instances with physiological data	
	Person 1	Person 2	Person 1	Person 2
None	648	433	319	433
Comment	79	9	13	9
Demand	18	14	4	14
Protest		5		5

Random oversampling is a simple method that randomly selects instances of the smaller classes and generates copies of them until all classes are balanced. Balancing was always done on training data only, so it could not happen that one copy of an instance would be in the training data and another in the test data.

Synthetic Minority Oversampling TEchnique (SMOTE) [1] generates new synthetic instances, and tends to perform better than random oversampling. It first randomly selects an instance from the minority class, then randomly selects a number of instances from among its five nearest neighbours (the number depending on the amount of oversampling required), and finally generates new instances as a linear combination of the original instances and each of the selected neighbours.

2.2 Feature Extraction

The features used for machine learning belong to three groups: (a) facial expressions and gestures; (b) vocalizations; and (c) physiological signals (see Figure 1). For each basic feature, we computed several derived features that were then used in our instances:

- AVG is the average of all values of a basic feature in the messages received within a (10-second) window.
- AVG>0 is the average of all the values that are greater than 0 (within a window).
- HIST is the histogram of values within a window; for each basic feature we obtain 10 histogram features.
- HISTD is the histogram density of values within a window; for each basic feature we obtain 10 histogram density features.

In addition to single derived features, the following combinations of derived features were considered: (a) AVG + HIST, (b) AVG + HISTD, (c) AVG > 0 + HIST, and (d) AVG > 0 + HISTD.

Note that physiological signals were not present in all data since Empatica was not used in all recordings. We tackled this issue by imputing physiological features with Decision tree regressor model when genuine physiological signals were not present. More precisely, one model was built for each physiological feature independently. To build the model, we used facial, gesture and vocalization data as well as annotated classes as features, while true values of the physiological features were used as target values.

2.3 Training Machine-Learning Models

Decision models were built with machine-learning algorithms and combined into ensemble models. Basic models were built with six diverse algorithms implemented in the Scikit-learn library [4]:

- Linear discriminant analysis (LDA)
- Gaussian naive Bayes (NB)
- Support vector machine (SVM)
- K-nearest neighbours (KNN)
- Decision tree (DT)
- Random forest (RF)

Among these models, RF typically performed well. However, RF is also sensitive to its hyperparameters. To further improve the accuracy, we tuned the following hyperparameters (evaluated values are in brackets):

- n_estimators [10, 20, 50, 100]
- max_features [0.1, 0.3, 0.6, 0.9]
- max_depth [2, 4, 6]
- min_samples_leaf [1, 2, 4]
- bootstrap [true, false]

Hyperparameter tuning was performed with grid search, stratified group k-fold cross-validation, and balanced accuracy. The tuned version of RF is hereafter referred to as RFO.

We implemented an additional model called the Expert system (ES). It consists of sets of rules that are derived from rule templates defined by the experts, i.e., caregivers. Each rule template has several conditions and the class into which it classifies. Each condition is defined with an attribute, i.e., the behaviour signal that is observed, the threshold that has to be met to trigger the condition, and the condition weight. In addition, a weight and a threshold are stored for each class value.

While the basic rules, i.e., rule templates were based on expert knowledge, we tuned the rules' parameters based on the annotated recordings. More precisely, the following parameter values were tuned: (a) class weights (for each class), (b) class thresholds (for each class), and (c) condition thresholds (for each condition in each rule). Parameter tuning was performed with the Differential Evolution algorithm [6].

2.4 Decision Fusion with Ensembles

The decisions of basic models were fused using an ensemble approach. More precisely, ensembles were built for each person and class independently by combining the best machine learning and expert models, and using various sets of features with and without physiological data.

Table 3: Ensembles

Name	Compulsory models	Number of best models	Voting type
b3	/	3	basic
b3	/	3	weighted
b2	/	2	weighted
rfo_b2	RFO	2	basic
rfo_b2_p	RFO	2	weighted
rfo_b1_p	RFO	1	weighted
es_b2	ES	2	basic
es_b2_p	ES	2	weighted
es_b1_p	ES	1	weighted
rfo_es_b1	RFO, ES	1	basic
rfo_es_b1_p	RFO, ES	1	weighted
rfo_es_p	RFO, ES	0	weighted

For each of the tested settings, the six machine-learning models in addition to ES and RFO were built (eight models in total). ES was preferred because it requires less data to be prepared for a new person, and RFO because it usually performed best. These models were sorted according to the balanced accuracies. Based on this, several ensembles were created as shown in Table 3. Each ensemble consists of compulsory models and additional models selected based on the balanced accuracies (b1 and b2 means best one or two models). The voting type was either basic (one model has one vote) or weighted (votes are weighted by the models' confidence).

3 EVALUATION OF PATTERN RECOGNITION APPROACH

To thoroughly test the possible configurations of the contextualized behaviour pattern recognition, we compared the accuracy of all combinations of:

- three class balancing methods (none, random oversampling and SMOTE)
- eight features sets (four single derived features and four combinations)
- with or without physiological signals
- 20 different models (six basic machine learning models, ES, RFO, and the 12 ensembles)

All this was done independently for inner states and communication attempts, and for persons 1 and 2, for a total of 1,280 experiments.

To evaluate the performance of the decision models, we used five-fold cross-validation: we split the data in five subsets, trained models on four and tested on the final one, and repeated this procedure five times with a different subset used for testing each time. We made sure that each example of inner state or communication attempt was only in one of the subsets, where one example means a continuous interval of a given state/attempt — this way we prevented the models from overfitting to specifics of a given situation.

Since our classes are imbalanced, we used balanced accuracy as the evaluation metric. Balanced accuracy is the average recall across

Table 4: Maximum balanced accuracy [%] of the different machine-learning models per person and class.

	Person 1 inner state	Person 1 comm. attempt	Person 2 inner state	Person 2 comm. attempt
ES	48.4	33.3	60.8	35.0
LDA	56.3	73.3	63.5	62.3
NB	53.5	58.0	64.8	44.8
SVM	60.2	57.0	67.1	48.8
KNN	57.9	51.1	59.1	44.8
DT	54.0	56.9	59.5	59.0
RF	49.3	52.4	62.5	35.0
RFO	59.8	70.5	65.5	60.8
b2_p	60.6	76.4	67.7	55.4
b3	62.1	72.5	69.3	53.6
b3_p	58.2	75.3	67.2	56.0
es_b1_p	59.3	74.6	68.0	0.0
es_b2	62.4	72.2	69.3	50.6
es_b2_p	59.2	76.4	66.9	0.0
rfo_b1_p	60.6	76.4	68.0	56.4
rfo_b2	62.1	72.5	67.3	53.6
rfo_b2_p	58.2	75.3	66.2	56.0
rfo_es_p	59.3	71.1	64.9	29.6
rfo_es_b1	60.6	72.2	67.0	50.6
rfo_es_b1_p	59.0	76.4	66.9	56.4

all the classes, and recall is the fraction of instances belonging to a class that are in fact recognized as such. Note that the balancing methods were used just on the training and not test data.

Table 4 shows the maximum balanced accuracies for individual persons and classes. Except in the last column, ensembles performed best. These best-performing ensembles include expert system as one of the models, i.e., es_b2 is the best model for person 1, inner state; es_b2_p is the best model for person 1, comm. attempt; and es_b2 is the best model for person 2, inner state. This suggests that it is beneficial to combine experts' domain knowledge in the form of expert system with artificial intelligence approaches that learn only from sensor data. On the other hand, among the single models, LDA and SVM performed best, closely followed by RFO.

Additional analysis was done regarding the various sets of features (see Table 5). For inner state, most feature sets performed similarly, although the more complex ones had a slight advantage. For communication attempt, the simpler feature sets proved better, probably because we had less training data, and so the models could not take advantage of a large number of features.

4 CONCLUSION

The behavior pattern recognition proved to be a challenging task. Since it processes data from recognizers (facial, gesture, etc.) that are themselves not perfect, it is affected by all their problems. In addition, the behaviours are sometimes difficult to annotate even for the caregivers, so there may be noise in the labels. The dataset is also limited, especially with respect to displeasure, since the caregivers try to avoid it. Furthermore, sometimes the person in the video is not detected, or the gestures and facial expressions are not recognized correctly. This may be due to the inherent difficulty of the task, or due to occlusion, as our experiment took place during normal care activities, so the caregivers moved around the persons with PIMD and placed them in diverse positions. In addition, not all behaviors that are informative for the caregivers can be recognized by the recognizers. Even without the previous problems, a person's behaviour is not always consistent; consequently, some behaviours are difficult to recognize even with correct inputs.

Table 5: Maximum balanced accuracy [%] of the different feature sets per person and class.

	Person 1 inner state	Person 1 comm. attempt	Person 2 inner state	Person 2 comm. attempt
AVG	58.4	76.4	66.8	44.8
AVG > 0	59.8	69.2	65.9	46.2
HIST	59.0	57.4	67.0	49.4
HISTD	59.3	57.0	68.0	62.3
AVG + HIST	60.2	57.6	67.1	45.3
AVG + HISTD	60.3	60.6	69.3	49.0
AVG > 0 + HIST	62.1	59.2	67.1	42.3
AVG > 0 + HISTD	62.4	57.8	68.5	48.8

The main task in our future work will more extensive evaluation of the proposed approach during INSENSION pilots. If time allows, we will collect additional data for the two persons in order to train the models on larger set of real-life situations. Since additional people will be involved in pilots, we will collect their data and build additional person-specific models.

ACKNOWLEDGMENTS

This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 780819.

REFERENCES

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal Of Artificial Intelligence Research* 16 (2002), 321–357.

[2] INSENSION Consortium. 2021. *INSENSION: Personalized intelligent platform enabling interaction with digital services to individuals with profound and multiple learning disabilities.* Retrieved June 9, 2021 from https://www.insension.eu/

[3] Michal Kosiedowski, Arkadiusz Radziuk, Piotr Szymaniak, Wojciech Kapsa, Tomasz Rajtar, Maciej Stroinski, Carmen Campomanes-Alvarez, B. Rosario Campomanes-Alvarez, Mitja Luštrek, Matej Cigale, Erik Dovgan, and Gašper Slapničar. 2020. On Applying Ambient Intelligence to Assist People with Profound Intellectual and Multiple Disabilities. In *Intelligent Systems and Applications*, Yaxin Bi, Rahul Bhatia, and Supriya Kapoor (Eds.). Springer International Publishing, Cham, 895–914.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, E. Duchesnay, and al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[5] Gašper Slapničar, Erik Dovgan, Pia Čuk, and Mitja Luštrek. 2019. Contact-Free Monitoring of Physiological Parameters in People With Profound Intellectual and Multiple Disabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops.*

[6] Rainer Storn and Kenneth Price. 1997. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. of Global Optimization* 11, 4 (Dec. 1997), 341–359. https://doi.org/10.1023/A:1008202821328