Understanding human behavioural patterns in naturalistic situations by explaining sequence model predictions

Bruno Michelot[1], Fabien Perrin[1]*, Stefan Duffner[2]*

[1] Centre de Recherche en Neurosciences de Lyon - INSERM U1028 - CNRS UMR 5292 - UCBL – UJM

[2] Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, F-69621 Villeurbanne, France

Abstract. This paper presents a new approach for detecting and analysing characteristic patterns of human behaviour from video, where the participants were filmed in situations involving different stimulations. After extracting features based on visual landmarks and facial action units, a set of binary classifiers is trained as a pretext task to discriminate pairs of situations. We propose a specific LSTM architecture that is designed for few and imbalanced training data using a *logsumexp* temporal aggregation function. Then, feature attributions based on Integrated Gradients are computed and analysed according to the different situations. By aggregating the classification explanations in several ways, we are able to highlight *which* features are most relevant for each situation and also the specific moments *when* they occur, which allows to interpret them further.

Keywords: Machine Learning · Human Behaviour · eXplainable Artificial Intelligence.

1 Introduction

The quantification and objectification of human behaviour in real-life situations are major topics of study in social neuroscience. Research has shown that our behaviour is influenced by the environmental context in which we find ourselves [15], and that these modifications can differ from those observed in traditional laboratory settings [25], where methodologies often overcontrol and constrain behaviour. Therefore, there is a need to develop new methods that can address these challenges. Recently, computer vision and machine learning approaches have gained increasing interest in these research fields for extracting and analyzing behavioural features. They enable the automatic and objective extraction of behaviour [2], even in situations that closely resemble real-life scenarios [7]. This approach allows to study the functional aspects of behaviour, which is central to the field of social neuroscience. However, these methods, in order to be effective, usually rely on supervised learning and are trained to do specific predictions based on labelled data which is very difficult to obtain.

In this context, we designed an experimental protocol and developed a new approach based on machine learning and eXplainable Artificial Intelligence (XAI) methods [3, 17] for the study of human behaviour in naturalistic situations. We recorded participants using a camera in various situations, including resting, listening to sounds or music, and listening to autobiographical or formal speeches from a person in front of them. We adapted existing computer vision algorithms [26, 9] to extract various low-level behavioural features for each video frame such as body and facial points, facial action units, and gaze direction. Subsequently, for each pair of situations, we trained a binary classifier based on a Long Short-Term Memory (LSTM) neural network architecture [12] specifically tailored for this type of data. Then, we applied Integrated Gradients [17] to obtain feature attributions of correctly classified examples. In that way, we were able to identify the behavioural characteristics associated with specific situations in which the participants were engaged and to examine how these characteristics were temporally modulated during the situation.

Our proposed approach proposes the following contributions:

- A method to identify characteristic behavioural patterns by first training a set of binary classifiers to discriminate pairs of situations as a pretext task and then computing feature attributions based on Integrated Gradients.
- A LSTM architecture that is designed for few and imbalanced training data using a *logsumexp* temporal aggregation function.
- Classification explanations aggregated in several ways and highlighting *which* features are most relevant for each situation and also the specific moments *when* they occur, which allows to interpret them further.

2 Related Work

Studies utilizing machine learning tools for behaviour analysis in neuroscience are relatively limited compared to more standard and commonly used methods on humans, such as motion sensors [23], eye trackers [4], and behavioural tasks that extrapolate behaviour [8]. The majority of studies using machine learning are focused on animal models, primarily due to the complexity involved in human models. Numerous computer vision models have been developed for various animal models, including tracking movements using centroids and ellipses estimation [10], colour pattern-based tracking [1], trajectory extrapolation [24], pose estimation and tracing [19, 11], 3D reconstruction [5], multiple target tracking [27], as well as studies utilizing software like DeepLabCut and DeepEthogram for prediction and classification based on extracted behavioural data [18, 6].

In human studies, behaviour analysis using machine learning benefits from the development of computer vision techniques, but it tends to be coarser. The focus is less on detailed behavioural analysis and more on predictive behavioural patterns associated with a disorder or prodromal stages of a disease, which lacks precision and detail. Examples include studies on schizophrenia [13], autism [16], Alzheimer's disease[16], suicidal behaviour [22], and sedentary behaviour [21] where the level of explainability is not satisfactory. Some recent studies have

adapted computer vision to investigate behaviour in social situations, particularly interpersonal synchrony [14], but they do not analyse behaviour holistically, despite the recognised importance of considering behaviour as a whole.

Thus, all these studies do not aim to demystify the behavioural foundations of these situations but rather utilise machine learning tools as a means to address specific research questions. Therefore, the current state of research in these fields lacks sufficient explainability techniques to provide a comprehensive understanding of how behaviour is modulated in everyday life situations.

3 Approach

3.1 Overall experimental setting

In our experiment, each participant was seated in a bed, in a semi-recumbent position and filmed with a camera from the front following a protocol of several repeated situations, each of one minute. There are 3×2 different types of situations : "stimulation", "rest", "social", with a self-referential and a control situation for each of the three. The "stimulation" situations consisted in presenting a sound to the participant while he or she was alone in the room: self-referential stimulation (SelfStim) was an excerpt of the participant's preferred music, and control stimulation (CtrlStim) was a white noise. The "rest" situations consisted in leaving the participant in silence after listening to the stimulation: self-referential rest (SelfRest) followed SelfStim while control rest (CtrlRest) followed CtrlStim. The "social" situations corresponded to a moment where the experimenter entered the room and recited a text: for the self-referential social situation (SelfSoc), this was a text related to the autobiographical contents associated with the music he or she listened just before (thus following SelfStim and SelfRest) while in the control social situation (CtrlSoc) this was a text about factual information (following CtrlStim and CtrlRest). The 6 situations were repeated 6 times for each participant.

From each frame of the recorded videos, we used OpenFace [26] to extract (a) 2D x/y coordinate points of the face (including contour, rotation, and facial elements), (b) the angle of gaze (in radians), and (c) the activation intensity of 17 Action Units (AU) (ranging from 0 to 5) based on the Facial Action Coding System (FACS) [20]. In addition, we used OpenPose [9] to detect the human "skeleton" via articulation points of the limbs. The 2D x/y coordinates of 15 points from the stem and the two upper limbs were extracted. Different preprocessing and normalisation operations are applied on these raw features in order to enhance their quality and make them usable for further analysis and machine learning (see Section 4.1 for more details).

3.2 Model and training

In order to extract relevant behavioural features, we train different classification models as a pretext task and then use the Integrated Gradient method (Sundararajan et al., 2017) to explain the classification. In theory, we could train a

multi-class model to classify the 6 situations explained in Sect. 3.1. However, there are many confusions between some of the classes, *e.g.* between SelfRest and CtrlRest or between SelfRest and CtrlStim, because the behaviour of some of the participants is almost identical in these situations leading to an overall accuracy below 20%. Therefore, we adopted an approach where we trained a binary classifier for each pair of situations resulting in 15 models. Naturally, the performance of each model increases because the task is simpler. But of greater interest is the fact that the IG feature attributions should be more pertinent and the increased flexibility to respond to specific questions w.r.t. to the behaviour in certain situations.

For the classification models, we propose to use a specific Long Short-Term Memory (LSTM) Neural Network that takes as input a sequence of extracted features from video (100) and is trained to predict the corresponding situation. The architecture is composed of 3 layers, a fully-connected (dense) layer of dimension 256, a LSTM layer of dimension 512 and a fully-connected layer with one output per class. The outputs of the hidden LSTM layer corresponding to all the time steps in the input sequence are aggregated and given to the final layer. The particularity of our architecture is that we use a *logsumexp* (LSE) function (Eq. 1) for this aggregation instead of a classical sum or another fully-connected layer as can be found in existing architectures.

$$LSE(x_1, \ldots, x_n) = \log \sum_{i=1}^{n} \exp(x_i) \tag{1}$$

This differentiable approximation of the max function can be seen as a type of smooth max pooling operation over the time steps in the hidden representations. This allows to cope with the imbalance in the data where crucial behavioural patterns occur relatively scarcely at specific moments of the sequence. More details on the training are given in Sect. 4.2.

3.3 Explaining model predictions

Once the models are trained, we apply each of them on a separate test set. Then we compute feature attributions for the correctly classified examples using IG, and we only take positive attributions.

To gain new insights into what behavioural cues are characteristic for a given situation, we computed the average attributions over all time steps of each situation and aggregated the attributions of several classifiers according to three behaviour aspects. The social effect ("Soc") was investigated through eight 2-2 comparisons: [SelfSoc or CtrlSoc] vs [SelfStim or SelfRest or CtrlStim or CtrlRest]. The stimulation effect ("Stim") was explored through four 2-2 comparisons: [SelfStim or CtrlStim] vs [SelfRest or CtrlRest]. The self-referential effect ("Self") was examined through three 2-2 comparisons: [SelfSoc] vs [CtrlSoc], [SelfStim vs CtrlStim], and [SelfRest vs CtrlRest].

To go further, we analysed the feature attributions over time in order to see *when* the relevant behaviour patterns occur within a sequence. The results are presented in Sect. 4.3

4 Experimental evaluation

4.1 Dataset

After feature extraction with OpenFace and OpenPose, we performed various preprocessing steps on the data (cleaning, smooting and normalisation). The OpenFace data underwent individual normalisation by subtracting the Action units average intensities for each participant from the data frame by frame. To eliminate biases related to size and position, the OpenPose data were normalised based on the participant's size and a relative reference point on the body. Additionally, x/y coordinate values were transformed into Euclidean distances. Movement was represented by calculating the difference between these distances values of a frame and the previous one. This resulted in the final dataset which comprises 16 participants, 662933 video frames in total and 100 features.

4.2 Training details

Each situation was split into overlapping sub-sequences of length 1000 with a step size of 60 during training and 10 during testing. We used the Adam optimization algorithm with beta1=0.99, beta2=0.999, weight decay=10e-4, a learning rate of 0.0007 and a batch size of 16. For each configuration, 3 training runs have been done, and early stopping is performed on a separate validation set.

4.3 Results

We computed the accuracy over all overlapping test sub-sequences. For final evaluation and analysis, we performed a 16-fold cross-validation, and, for each split, kept out the data of 1 participant for test. First, we evaluated the effectiveness of the LSTM architecture, in particular the *logsumexp* aggregation function (Eq. 1), with models trained on 3 classes: 1) SelfStim + CtrlStim 2) SelfRest + CtrlRest 3) SelfSoc + CtrlSoc. We compared the *logsumexp* aggregation to a simple sum and the max function. The results are shown in Table 1 and clearly show the superiority of our proposed model in this context.

Aggregation type	Accuracy
max	0.667 ± 0.035
sum	0.653 ± 0.032
logsumexp	$\mathbf{0.855 \pm 0.038}$

Table 1. Test accuracy for LSTM models with different types of aggregation functions before the final layer (Mean + SEM).

As mentioned in Sect. 3.3, we then computed the IG feature attributions for each feature and each fold (participant) on the binary classifiers trained on the

15 possible combinations of situations. These values were ranked from highest to lowest for each participant and the average ranking was calculated for each classifier. In parallel, we regrouped the behavioural features into clusters based on specific facial areas and feature types, *i.e.* AU - Eyes, AU – Mouth, AU – Nose, Points – Eyes, Points – Mouth, Points – Nose, Points – Limbs, Points – Head Contour, and Gaze. For each classifier, the 5 most important features were retained. If a feature cluster had at least one component present in the top 5, a value of 1 was assigned to that cluster. Otherwise, a value of 0 was assigned to the cluster. Finally, the average of these values (1 and 0) was calculated within the "Soc", "Stim", and "Self" groups. Figure 1 illustrates that the key behavioral

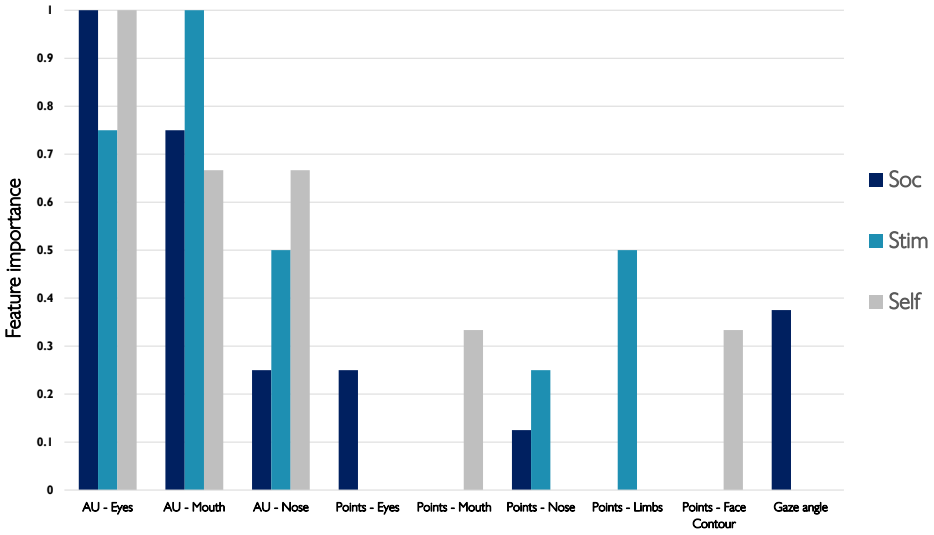

Fig. 1. Relevant groups of behavioural features for the classification of the different situation groups "Soc", "Stim" and "Self".

features for classifying different types of situations predominantly focused on facial features, specifically the various groups of AUs. The AUs related to the eyes had a score of 1 for the Soc and Self groups, while the AUs related to the mouth had a score of 1 for the Stim group. The AUs related to the nose were also utilized, particularly for the Self group (score of 0.666). In addition to AUs, each group exhibited distinct behavior characteristics that were crucial for classification. For instance, to specifically differentiate social situations ("Soc"), the algorithm also relied on eye-related points (score of 0.25) or gaze angle (0.375) in some classifications. In the case of stimulation situations ("Stim"), limb-related points (score of 0.5) were specifically decisive for classification, whereas points related to the mouth and head contour (scores of 0.33 for both) were specific in distinguishing emotional situations ("Self") from others.

Finally, Fig. 2 shows the temporal evolution of the different feature groups for one participant in different situations. For better visibility, we have aggre-

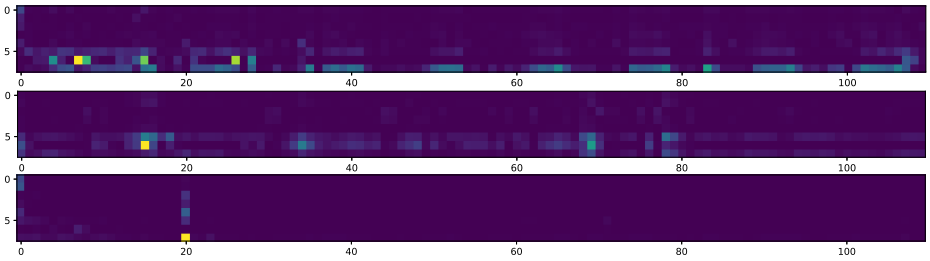

Fig. 2. Feature attributions for one participant over time for different situation. Top: SelfStim situation explained w.r.t. SelfRest. Middle: SelfSoc explained w.r.t. SelfStim. Bottom: SelfStim explained w.r.t. CtrlStim. Each line in each image corresponds to the average of a group of features, from top to bottom: points face contour, eyes, nose, mouth, AU eyes, AU nose, AU mouth. (1 pixel=0.5s).

gated several time steps, such that one pixels corresponds to 0.5s. The relevant moments are clearly highlighted (in brighter colours), and correspond mostly to AUs.

5 Conclusion

We presented an effective approach for analysing human behaviour in naturalistic situations using machine learning and XAI. Our method trains one-vs-one binary classifiers based on an LSTM model with a specific architecture and then uses IG feature attributions that are subsequently ranked and grouped according to different situations in order to highlight the behavioural features that are characteristic for a given situation. To go further, we analysed the temporal evolution of feature attributions in order to know when different characteristic behavioural patterns occur within a sequence. The obtained results are very encouraging and open new ways of getting insights into behaviour patterns for multi-disciplinary research.

References

1. A. Gal, J. Saragosti, D.J.C.K.: anTraX: high throughput video tracking of color-tagged insects. preprint, Neuroscience (2020)
2. A. Voulodimos, N. Doulamis, A.D., Protopapadakis, E.: Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience (2018)
3. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–-52160 (2018)

4. B. T. Carter, S.G.L.: Best practices in eye tracking research. International Journal on Psychophysiology **155**, 49–62 (2020)
5. Badger, M., Wang, Y., Modh, A., Perkes, A., Kolotouros, N., Pfrommer, B.G., Schmidt, M.F., Daniilidis, K.: 3D Bird Reconstruction: a Dataset, Model, and Shape Recovery from a Single View. http://arxiv.org/abs/2008.06133 (2020)
6. Bohnslav, J.P., Wimalasena, N.K., Clausing, K.J., Dai, Y.Y., Yarmolinsky, D.A., Cruz, T., Kashlan, A.D., Chiappe, M.E., Orefice, L.L., Woolf, C.J., Harvey, C.D.: Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. eLife **10, e63377** (2021)
7. C. Ionescu, D. Papava, V.O., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
8. Draheim, C., Tsukahara, J.S., Martin, J.D., Mashburn, C.A., Engle, R.W.: A toolbox approach to improving the measurement of attention control. Journal of Experimental Psychology General **150**, 242–275 (2021)
9. G. Cao, T. Hidalgo, S.E.W.S., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2021)
10. Geuther, B.Q., Deats, S.P., Fox, K.J., Murray, S.A., Braun, R.E., White, J.K., Chesler, E.J., Lutz, C.M., Kumar, V.: Robust mouse tracking in complex environments using neural networks. Commun. Biol. **2**(124) (2019)
11. Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D.: Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. eLife **8, e47994** (2019)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**, 1735–80 (1997)
13. Jiang, Z., Luskus, M., Seyedi, S., Griner, E.L., Rad, A.B., Clifford, G.D., Boazak, M., Cotes, R.O.: Utilizing computer vision for facial behavior analysis in schizophrenia studies: A systematic review. PLoS ONE **17, e0266828** (2022)
14. Koul, A., Ahmar, D., Iannetti, G.D., Novembre, G.: Interpersonal synchronization of spontaneously generated body movements. Science **26, 106104** (2023)
15. Krishnan-Barman, S., Forbes, P.A.G., de C. Hamilton, A.F.: How can the study of action kinematics inform our understanding of human social interaction? Neuropsychologia **105**, 101–-110 (2017)
16. Li, X., Fan, F., Chen, X., Li, J., Ning, L., Lin, K., Chen, Z., Qin, Z., Yeung, A.S., Li, X., Wang, L., So, K.F.: Computer vision for brain disorders based primarily on ocular responses. Front. Neurol. **12, 584270** (2021)
17. M. Sundararajan, A.T., Yan, Q.: Axiomatic attribution for deep networks. arXiv (2023)
18. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. **21**, 1281–1289 (2018)
19. Mathisi, M.W., Mathis, A.: Deep learning tools for the measurement of animal behavior in neuroscience. Curr. Opin. Neurobiol. **60**, 1–11 (2020)
20. P., E., W., F.: Facial action coding system: A technique for the measurement of facial movement. Palo Alto: Consulting Psychologists Press (1978)
21. Papathomas, E., Triantafyllidis, A., Mastoras, R.E., Giakoumis, D., Votis, K., Tzovaras, D.: A machine learning approach for prediction of sedentary behavior based on daily step counts. In: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 390–394 (2021)

22. Paska, A.V., Kouter, K.: Machine learning as the new approach in understanding biomarkers of suicidal behavior. Bosn. J. Basic Med. Sci. (2020)
23. Preece, S.J., Goulermas, J.Y., Kenney, L.P.J., Howard, D., Meijer, K., Crompton, R.: Activity identification using body-mounted sensors–a review of classification techniques. Physiological Measurements **30**, R1–33 (2009)
24. Rodriguez, A., Zhang, H., Klaminder, J., Brodin, T., Andersson, P.L., Andersson, M.: Toxtrac: A fast and robust software for tracking organisms. Methods Ecol. Evol. **9**, 460–464 (2018)
25. S. L. Rogers, C. P. Speelman, O.G., Longmuir, M.: Using dual eye tracking to uncover personal gaze patterns during social interaction. Sci Rep. **8**(1), 4271 (2018)
26. T. Baltrusaitis, P.R., Morency, L.P.: Openface: An open source facial behavior analysis toolkit. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2016)
27. V. H. Sridhar, D. G. Roche, S.G.: Tracktor: Image-based automated tracking of animal movement and behaviour. Methods Ecol. Evol. **10**, 815–820 (2019)