Journal of Neural Engineering

Neural correlates of face perception modeled with a convolutional recurrent neural network

Journal of Neural Engineering

PAPER

Neural correlates of face perception modeled with a convolutional recurrent neural network

Jamie A O'Reilly[1], **Jordan Wehrman**[2,*], **Aaron Carey**[3], **Jennifer Bedwin**[3], **Thomas Hourn**[3], **Fawad Asadi**[4] and **Paul F Sowman**[3]

[1] School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand
[2] Mind and Brain Institute, Department of Psychology, University of Sydney, New South Wales 2006, Australia
[3] School of Psychological Sciences, Macquarie University, New South Wales 2019, Australia
[4] College of Biomedical Engineering, Rangsit University, Pathum Thani 12000, Thailand
[*] Author to whom any correspondence should be addressed.

E-mail: jordan.wehrman@sydney.edu.au

Keywords: neural correlates, face perception, recurrent neural networks, convolutional neural networks, EEG modelling, N170

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
16 January 2023

REVISED
19 February 2023

ACCEPTED FOR PUBLICATION
10 March 2023

PUBLISHED
3 April 2023

Abstract

Objective. Event-related potential (ERP) sensitivity to faces is predominantly characterized by an N170 peak that has greater amplitude and shorter latency when elicited by human faces than images of other objects. We aimed to develop a computational model of visual ERP generation to study this phenomenon which consisted of a three-dimensional convolutional neural network (CNN) connected to a recurrent neural network (RNN). *Approach.* The CNN provided image representation learning, complimenting sequence learning of the RNN for modeling visually-evoked potentials. We used open-access data from ERP Compendium of Open Resources and Experiments (40 subjects) to develop the model, generated synthetic images for simulating experiments with a generative adversarial network, then collected additional data (16 subjects) to validate predictions of these simulations. For modeling, visual stimuli presented during ERP experiments were represented as sequences of images (time x pixels). These were provided as inputs to the model. By filtering and pooling over spatial dimensions, the CNN transformed these inputs into sequences of vectors that were passed to the RNN. The ERP waveforms evoked by visual stimuli were provided to the RNN as labels for supervised learning. The whole model was trained end-to-end using data from the open-access dataset to reproduce ERP waveforms evoked by visual events. *Main results.* Cross-validation model outputs strongly correlated with open-access ($r = 0.98$) and validation study data ($r = 0.78$). Open-access and validation study data correlated similarly ($r = 0.81$). Some aspects of model behavior were consistent with neural recordings while others were not, suggesting promising albeit limited capacity for modeling the neurophysiology of face-sensitive ERP generation. *Significance.* The approach developed in this work is potentially of significant value for visual neuroscience research, where it may be adapted for multiple contexts to study computational relationships between visual stimuli and evoked neural activity.

1. Introduction

Visual perception of objects is a combined function of multiple neural systems and cognitive processes. Simply put, the visual system processes visual information received by the eyes, while the cognitive system organizes and interprets this information. These two systems work together to allow us to perceive and understand the objects and scenes we encounter in the world around us. Faces represent a special object category in human perception. Face perception has evolved in humans to be highly specialized and efficient [1], reflecting the central role that faces play in social interactions and communication. The perception of faces is the primary information source for the recognition and identification of specific individuals, is critical for interpreting expressions and emotions, and underpins the extraction

of important social cues; hence, impaired visual or visual-cognitive perceptual systems can severely challenge social engagement. In light of this, investigating the relationship between face perception and neural activity may help us to understand putative disruptions to neural systems in clinical conditions such as autism [2], congenital or acquired prosopagnosia [3], and schizophrenia [4], which often involve atypical face perception as measured by behavioral and neurophysiological measures.

A consistent finding from event-related potential (ERP) studies concerning face perception is sensitivity of visually evoked N170 to images of human faces [5–8]. This ERP peak occurs earlier in response to human faces than other objects and typically has greater amplitude at right hemisphere occipitotemporal electrode sites associated with face processing specialization [9, 10]. There remains considerable debate in the literature about whether the N170 component is specific to faces or rather represents a response to highly familiar stimuli—of which faces invariably are [6–10]. One argument often cited in support of the specificity view is the observation that familiar but non-face objects tend to elicit an N170 later in time than faces do. This suggests that the N170 may be more closely tied to face processing than to general familiarity with a stimulus, and supports the idea that it reflects a specific brain process devoted to face perception. However, these two explanations are not necessarily mutually exclusive, and scalp-recorded ERP signals may well reflect overlapping representations of both face specificity and more general object familiarity.

This argument requires methodological improvements to resolve. Promising, data-driven signal extraction techniques have recently become tenable through the proliferation of machine learning methods and high-performance computing. In particular, deep artificial neural networks capable of modeling complex nonlinear functions relating inputs to outputs can potentially yield valuable insights from neurophysiological data [11, 12]. For example, convolutional neural network (CNN) models have been used to investigate the computational principles underlying visual perception [13]. These are machine learning models specialized in visual processing tasks such as object recognition. Operations performed by CNNs are considered mechanistically analogous to neurobiological implementations of vision, thus providing useful models for studying these systems [13]. While CNNs are adept at image processing, standard architectures lack the required memory elements to handle sequential, time-varying data [13]. Recurrent neural networks (RNNs) solve this problem with recurrent connections that provide information from previous time-steps, thereby enabling the network to model sequences. These are widely used in natural language processing, and have recently been applied to model ERP waveforms [14–17]. Examples of RNN use in evoked response modeling include simulating ERPs to different stimuli [14, 15], estimating changes in ERP morphology between states of consciousness [17], and exploring the computational processes of auditory-evoked potential generation [14–16]. To this end, the differences in N170 characteristics discussed above make a suitable test case for the use of CNN-RNN combined analysis of ERPs. Not only can we analyze how these neural networks make sense of the data, helping to resolve disputes such as the familiarity-face specificity debate mentioned above, but also by training with data from one condition (such as upright faces) and testing network outputs to alternative inputs (such as inverted faces), we can examine how neural networks can or cannot simulate established effects. By exploring this possibility in the current article, we can also comment on how such effects are 'cognitive' in nature rather than occurring 'mathematically' as part of the neural architecture.

These new analytical techniques based on machine learning algorithms can be applied to ERP data from face perception studies to help interpret the neurophysiology elicited by viewing different images. In this article we propose a computational model of human visually-evoked responses by combining the image processing capabilities of a CNN with the sequence learning functions of an RNN. We then fitted it to normative data from a classic N170 experiment. Variants of this design have been used in other contexts. For example, Shi *et al* used a CNN-RNN model for action recognition in video clips [18], and Xu *et al* used a similar architecture for epileptic seizure detection [19]. Our approach is different insofar as we trained our model to reproduce human ERP waveforms in response to sequences of input images that were presented to subjects in a face perception study. Behavior of the model was then analyzed to investigate its correspondence with theories about ERP sensitivity to face and non-face images. Grill-Spector *et al* (2018) [20] highlighted the potential for using deep neural network architectures to model the ventral face processing network in humans. However, deep-learning approaches for face perception (e.g. Google's 'Facenet') have mainly concentrated on computer vision detection and identification of faces rather than studying human neurophysiology.

The main goal of this study was to develop and cross-validate a CNN-RNN model of visually-evoked potentials for studying the neural signature of face perception in ERP waveforms. To achieve this, we used three datasets. Firstly, data from the ERP Compendium of Open Resources and Experiments CORE (EC) dataset was used to train a model to produce ERPs based on face, car and scrambled images. This provided basic proof of concept for the method. We then generated synthetic images to examine the ability of the trained CNN-RNN model to produce reasonable ERPs outside of the training dataset. In addition,

we also inverted the images to check how the CNN-RNN would generalize to common manipulations of face-car datasets. Finally, we generated a new dataset by recording from naïve participants. This allowed us to compare with the ERP CORE dataset and evaluate the performance of the CNN-RNN predictions of manipulated images.

2. Methods

The methods are illustrated graphically with a flowchart in supplementary figure S1. The author's have confirmed that any identifiable participants in this study have given their consent for publication.

2.1. Data

2.1.1. ERP CORE dataset

The face N170 dataset was downloaded from the ERP CORE (Compendium of Open Resources and Experiments) repository [21]. For brevity this dataset will be abbreviated to EC. This consisted of electroencephalography (EEG) recorded from 40 healthy adult human subjects viewing a pseudo-random sequence of face (F) and car (C) images and their phase-scrambled counterparts (SF and SC) [6]. There were 40 images in each category (i.e. 160 images in total). All images were presented twice; once in the first half and once in the second half of the experiment. In other words, once all images had been presented, the order of images was scrambled again and then presented again for the second half of the experiment. Signals were band-pass filtered from 0.1 to 20 Hz then resampled to 100 Hz before re-referencing to the average of 33 electrodes [21]. Channels PO7 and PO8 were then selected for further analysis. Epochs were extracted from 0.2 s before to 0.8 s after each image presentation. Images were shown for 0.3 s in the experiment. Baseline correction was applied by subtracting the average pre-stimulus value from the whole epoch. The grand-average ERP waveform from each stimulus was produced by averaging across subjects. This resulted in 160 ERPs measured from two channels that were 101 samples in length. These provided labels for optimizing the model with supervised learning [16].

2.1.2. Synthetic dataset

Artificial images were synthesized using generative adversarial network (GANs) [22] to augment the real image data with nonidentical synthetic images from an overlapping distribution of pixel values and configurations. These synthetic images were subsequently used in simulated experiments with the trained CNN-RNN model. Pre-trained StyleGAN2 models were fine-tuned with active discriminator augmentation [23] to generate images comparable with EC stimulus images. To prepare training data for these GANs, real images were resized to 512 px high with bi-cubic interpolation then padded to 512 px wide with picture elements matching the background gray value of 205 from the 8-bit bitmap format images. Training images were saved in 8-bit portable network graphics format. Two training sets were produced: one with 40 face images and another with 40 car images, each used to fine-tune a separate GAN. Both of the GANs had been pre-trained on the Flickr-Faces-HQ dataset [24]. We explored the same GAN architecture pre-trained on car images [24] to generate car images, but found that this was outperformed by the GAN pre-trained on face images. This is considered to be due to pre-training face images being higher resolution (1024 by 1024 px) than pre-training car images (512 by 384 px). Both GANs were implemented on a computer system with a graphics processing unit (Titan RTX 24 GB, Nvidia; Santa Clara, CA, USA). Batch size for each training step was 32. Fréchet Inception Distance (FID) [25] was computed every fourth step and training stopped if no improvement in FID was observed for 160 consecutive steps. The face-generating GAN was trained for 8900 epochs, achieving minimum FID score of 21.67; the car-generating GAN trained for 3300 epochs, achieving FID score of 50.41.

One thousand artificial face and car images were produced by the GANs after training. Truncation for StyleGAN2 was set to 0.65; the range of values this parameter can take is from 0 to 1, and it controls a trade-off between maximum fidelity (0) and diversity (1) of generated images. The resulting images were cropped and resized to match the format of original EC stimulus images. By visual inspection and FID scores these synthetic images were deemed to be comparable with but not identical to images in their respective training sets. Some synthetic face images appeared to be slightly different images of persons from the training dataset. Other synthetic face images can be described as combinations of two or more persons found in the training dataset. Synthetic car images generally looked genuine, although the interiors were relatively poorly constructed and seemed to depict a row of three front seats. Despite these and some other minor limitations, such as asymmetrical front light designs, synthetic car images were deemed adequate. All of the synthetic images are available from [https://osf.io/y36e5/?view_only=5ec917a504a1492ab0b69f4d2500f485]. Importantly, none of the artificial images were identical to any of the real images found in the training sets. These generated images were also phase-scrambled to produce a complete synthetic image dataset containing one-thousand images of each stimulus category.

2.1.3. Validation dataset

An experiment was designed to collect out-of-sample data to compare with model outputs and EC data. This design was reviewed and approved by the Macquarie University (MQ) ethics committee. This

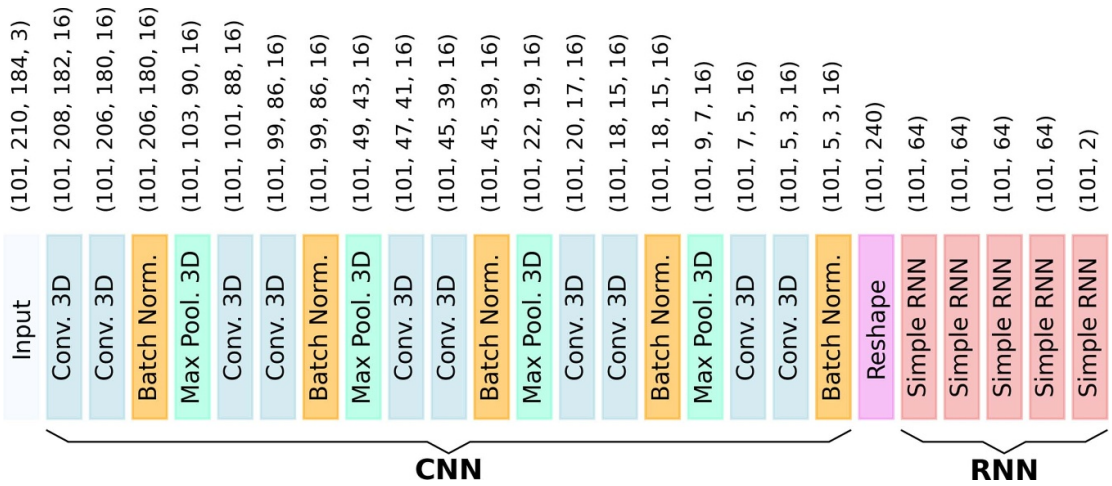

Figure 1. Model architecture. CNN and RNN sections are annotated in bold. Sequentially connected layers are illustrated with colored blocks. The shape of output tensors from each layer is noted above each colored block. Convolutional layers decreased the size of spatial dimensions by two picture elements because of the filter size and absence of zero-padding. Max pooling layers decreased the spatial scale by a factor of two, rounded down for odd-numbered sizes. Temporal structure (101 time-samples) of the data is maintained throughout the network.

validation dataset is abbreviated to MQ. Identical images to those used in EC [6] were shown to 16 healthy adult human subjects. Face (F), car (C), scrambled face (SF) and scrambled car (SC) images were presented upright and inverted (upside-down) in three different stimulus blocks counterbalanced across subjects. The first block included F, C, SF and SC conditions, equivalent to the EC sequence. The second block contained upright and inverted F and C images. The third block comprised inverted F, C, SF and SC images. Each block included 160 images played twice. Presentation® (20.2, Neurobehavioral Systems Inc.; Berkeley, CA, USA) software was used to present these sequences with synchronized, event-coded time-stamps. These event codes were delayed to account for monitor refresh rate. Electroencephalography signals were recorded using a BioSemi® Active Two (BioSemi Systems; Heerlen, the Netherlands) with an equivalent montage to the EC dataset. Filtering, resampling, re-referencing, channel selection, epoch segmentation, baseline correction and grand-average ERP calculation were performed as described in section 2.2.1. above. This data is available at: https://osf.io/y36e5/

2.2. Modeling

2.2.1. Architecture and training

The model design can be subdivided into CNN and RNN sections illustrated in figure 1. Input tensors were sized 101 (time-samples) by 210 (image height) by 184 (image width) by 3 (color channels). These fed into the first of ten three-dimensional convolutional layers. Each convolutional layer had 16 filters sized 1 (time-sample) by 3 (height) by 3 (width), effectively filtering over spatial dimensions of the image data while preserving its temporal structure. Filter activations were rectified and no zero padding was applied. Each pair of consecutive convolutional layers was followed by a batch normalization layer [26]. The first four batch normalization layers were followed by three-dimensional max pooling layers [27] with kernel size of 1 (time-sample) by 2 (height) by 2 (width), causing tensors to compress over spatial dimensions. Outputs from the fifth batch normalization layer were reshaped to 101 (time-samples) by 240 (filter activation maps), effectively a time-sequence of vectors representing the input image sequence.

Sequences of vector representations from the CNN section were fed into the RNN section. The first four simple recurrent layers had 64 units with the rectified linear unit activation function applied. The fifth recurrent layer was the model output. It had two units, one for reproducing ERP data from channel PO7 and the other for PO8. The model was trained end-to-end. Input data represented visual stimuli presented to humans in the face N170 experiment, and output labels were derived from corresponding ERP waveforms. The ERP epoch spanned −0.2 s to 0.8 s around each stimulus onset time, sampled at 100 Hz. Stimulus images were shown for 0.3 s, but otherwise the screen was gray. This scenario was represented with arrays containing 101 images: the first twenty were gray (−0.2 s to 0.0 s), the next thirty were copies of the image presented to subjects in the experiment (0.0–0.3 s), and the remaining fifty-one (0.3–0.8 s) were gray. These were paired with output labels obtained by multiplying corresponding ERPs by 1 million to normalize their units (ERPs are typically in microvolts). The model was thereby optimized to produce outputs matching ERP waveforms in response to sequences of images shown in the experiment. Training was performed for 1000 epochs with a batch size of 8. The adaptive moment

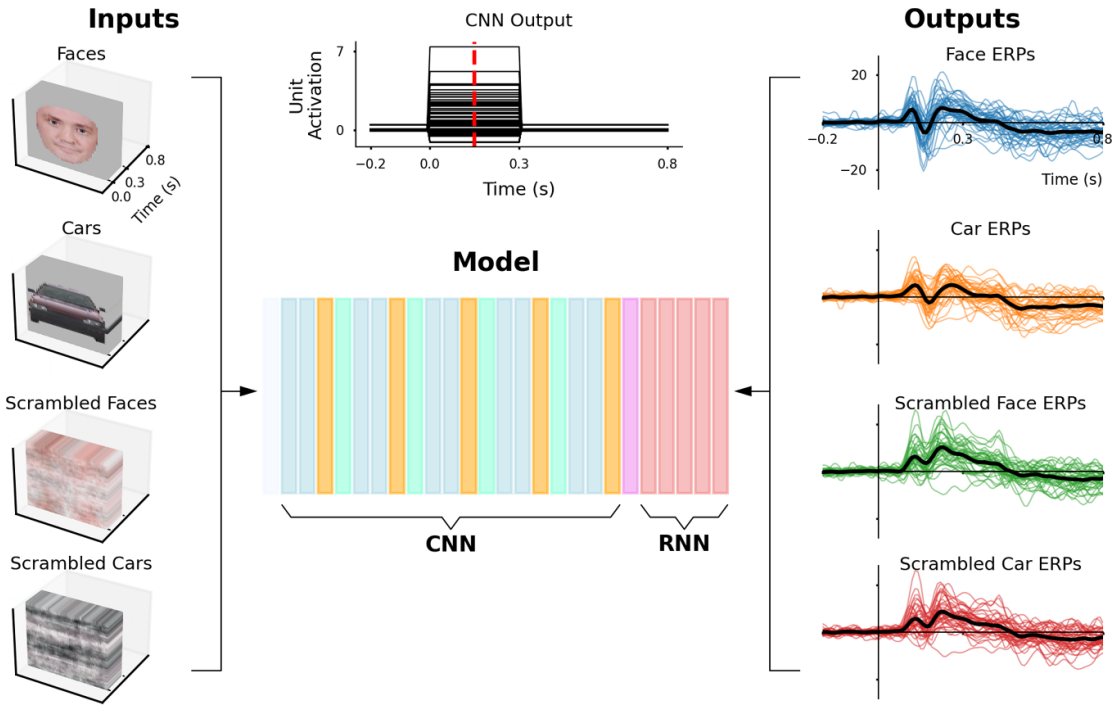

Figure 2. Supervised learning paradigm. Inputs on the left-hand side were three-dimensional representations of image time-sequences presented in the visual ERP experiment. Four stimulus categories were presented: faces, cars, scrambled faces and scrambled cars, each comprising 40 unique images. The computational model is illustrated conceptually in the center. It consisted of a three-dimensional CNN connected to a RNN. The CNN transformed each image of the input time-sequence into a vector representation. An example of CNN outputs for one input image sequence is shown at the top, with a red dashed line transecting the vector representation during stimulus image presentation from 0.0 to 0.3 s. The model was trained end-to-end, with output labels provided from ERP waveforms recorded from a cohort of human subjects as they viewed the same stimuli depicted by the inputs. Grand-average ERPs are identified with thick black traces.

estimation optimizer was used with default parameters (0.001 learning rate, 0.9 beta-1, 0.99 beta-2) and mean-squared error loss. Input arrays, computational model, CNN vector outputs, and output labels are illustrated in figure 2.

The model was trained with 160 image-ERP pairs from EC. The trained model was used to produce 4×4000 synthetic image-ERP pairs from simulated experiments with three manipulations applied to synthetic images (inverted, low-pass filtered (LP), and high-pass filtered (HP)). The validation dataset comprised 2×160 image-ERP pairs from upright and inverted images that were compared with EC ERPs and synthetic ERPs from simulations. During training, the model was optimized to produce outputs that resemble ERP waveforms elicited by corresponding image sequences. Therefore, the primary outputs of the model are generated ERPs (figure 3). After training, image vector representations from the CNN (figure 4) and patterns of hidden unit activations (figure 6) can be extracted, so they may be considered secondary outputs of the model.

2.2.2. Cross validation

The training dataset from EC was used for cross-validation (CV). The model architecture was evaluated with ten-fold cross validation. In each fold, four different pairs of input arrays and output labels from each stimulus category (F, C, SF and SF) were set aside to be used as validation data. A fresh model was instantiated and trained using the remaining 144 input-label pairs. After training, models were used to predict ERP waveforms in response to held-out validation data, producing a complete set of 160 CV ERPs over ten folds. Overall performance was evaluated by comparing how well these model outputs correlated with real ERPs from EC and MQ datasets, displayed on figure 3. After CV, a final model was trained using all of the available EC data. The behavior of this final model was then analyzed to investigate how it generates outputs replicating ERP waveforms in response to sequences of visual input.

2.2.3. Model analysis

2.2.3.1. Principal component analysis of image vector representations

Principal component analysis was used to reduce the dimensionality of CNN vector representations from 240 to 2 for visualization (i.e. we ran a principal component analysis in which the data was reduced to 2 components). The vector representing each stimulus image was retrieved from a single time-slice of the CNN output, taken from a time-point when the stimulus image was shown, illustrated by the dashed

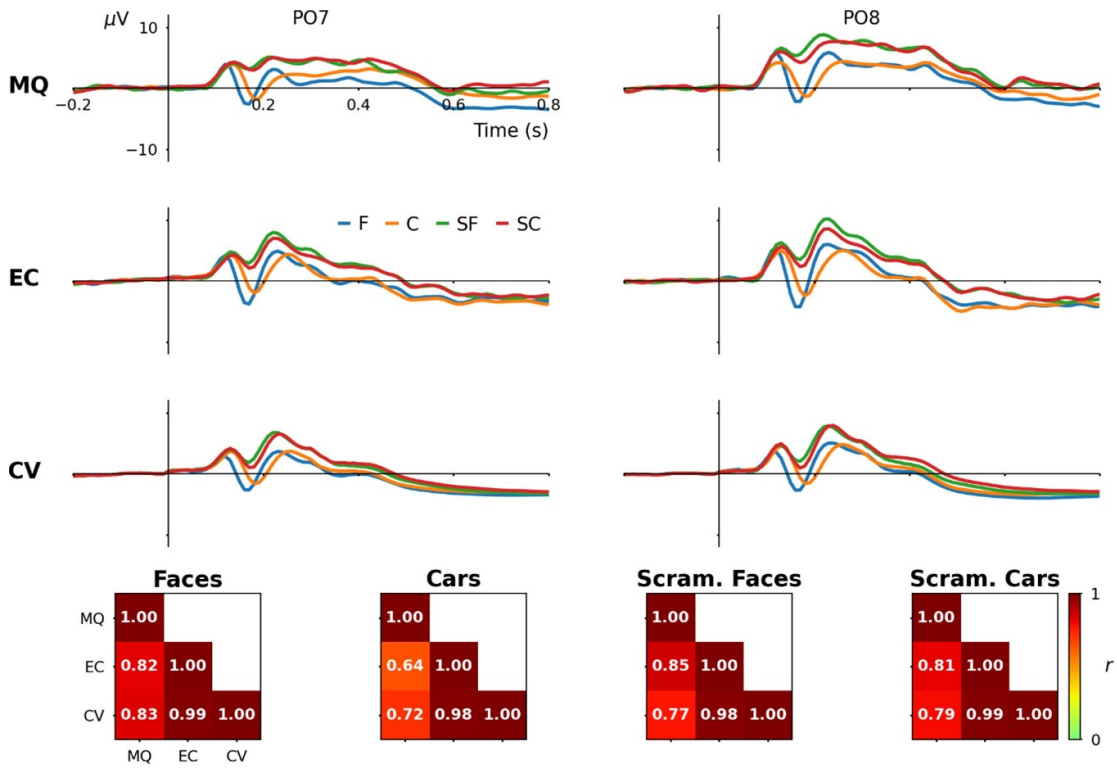

Figure 3. Grand-average ERP waveforms from two human cohorts and computational model. The first row displays data from the MQ validation study ($n = 16$); the second row shows data from the EC dataset ($n = 40$); the third row presents outputs from model CV cases ($n = 40$). The bottom row depicts correlation (r) across the three datasets for face (F), car (C), scrambled face (SF) and scrambled car (SC) ERP waveforms. For all correlations, $p < 0.001$.

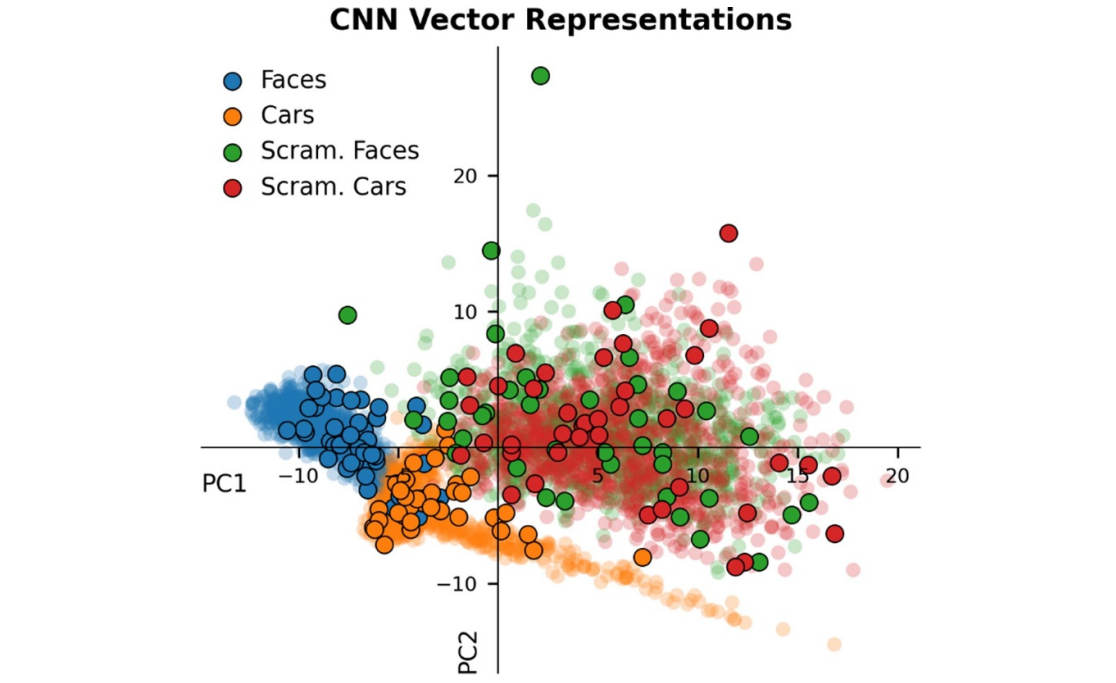

Figure 4. Principal component analysis of vector representations learned by the CNN. Each vector representation of 240 elements was transformed into two principal components for visualization. Real image data points ($n = 40$ for each image category) are shown by solid markers with black outlines. Markers representing synthetic image data points ($n = 1000$ for each category) are slightly transparent. Notably, the separation of faces and cars from scrambled images demonstrates the separation of 'real world' images from scrambled images in the representations learned by the CNN.

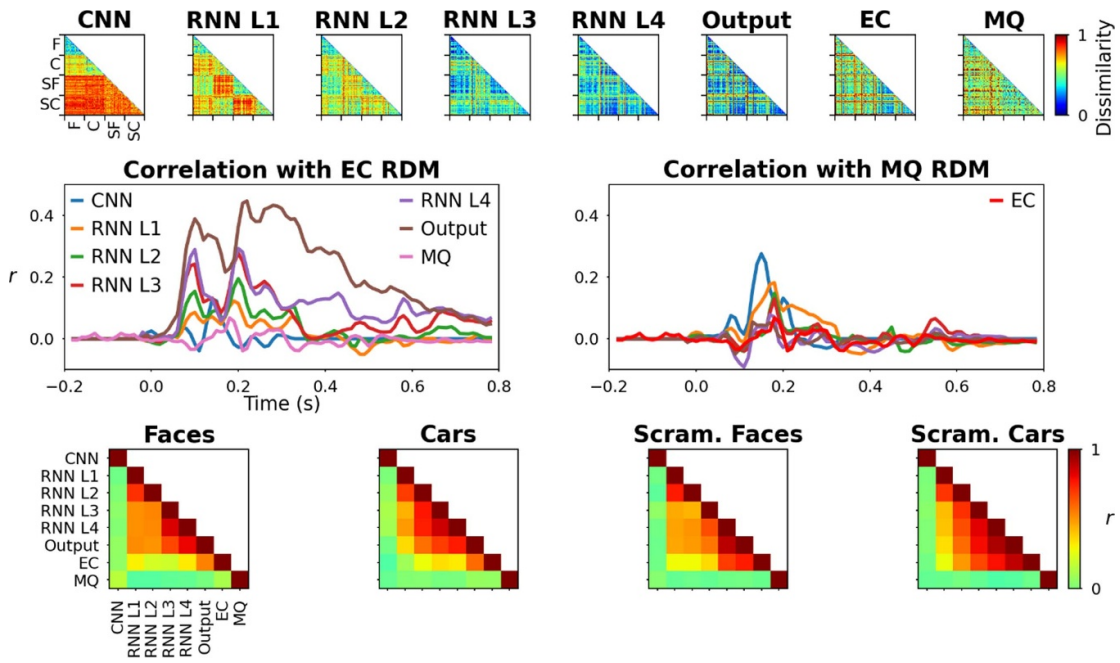

Figure 5. Representational similarity analysis. The top row shows RDMs from CNN outputs, four RNN hidden layers (L1–L4), model output, EC ERPs and validation study (MQ) ERPs. These 160 by 160 matrices were computed by comparing 40 samples of four stimulus categories: faces (F), cars (C), and scrambled faces (SF) and cars (SC). The middle-left panel shows correlation between model and MQ with EC RDMs calculated from a five-sample sliding window with step size of one. The middle-right panel displays similar time-wise correlation between model and EC RDMs with the MQ RDMs. The bottom row displays correlation between RDMs in the top row by image category.

line in the upper middle panel of figure 2. These vector representations were generated by the final model that had been trained on the whole EC dataset. The singular value decomposition transform was determined from vector representations of EC stimuli. Both EC and synthetic image vector representations were then transformed into two principal components for visual comparison. These data are plotted in figure 4.

2.2.3.2. Classification analysis of image vector representations

To further evaluate vector representations produced by the CNN section, a logistic regression classifier was trained on vector representations from 160 real EC images and then tested on those from 4000 upright synthetic images. This was treated as a one-versus-rest multi-class classification problem, with four classes of images represented by the vectors. Comparing classification performance across the four categories of images provides information about how robustly these are represented by the 240 element vectors. For example, classes with high accuracy occupy a designated region in multidimensional space and have distinctive vector representations, whereas classes with low accuracy overlap in vector space and have less distinctive vector representations. Within the described end-to-end training paradigm, robustness of these vector representations will influence how well model outputs match ERP labels.

2.2.3.3. Representational similarity analysis of model behavior and ERPs

Representational similarity analysis (RSA) [28] was applied to examine final model behavior relative to EC and MQ ERP waveforms. This is shown in figure 5. Activations were obtained from the model at multiple levels: CNN output, RNN layer 1–4, and model output. Representational dissimilarity matrices (RDMs) were computed across the full epoch and also across five-sample sliding windows with step size of one. Correlations among both sets of ERP data and model activation RDMs were calculated. The RSA approach allows us to compare how model activations and neurophysiological signals vary in response to the same set of stimuli.

2.2.3.4. Analysis of model RNN hidden units associated with face and car images

Hidden unit activations from the RNN section before the output were categorized as being associated mostly with face images, car images, or neither, according to the stimuli that elicited their maximal response. This was determined by integrating hidden unit activations over time and classifying them according to the stimulus category that maximized this. Three class separation was justified by the patterns of ERP waveforms and image vector representations that suggested overlap between the two types of scrambled images. Therefore, responses to both sets of scrambled images were grouped together and this analysis concentrated on differences between

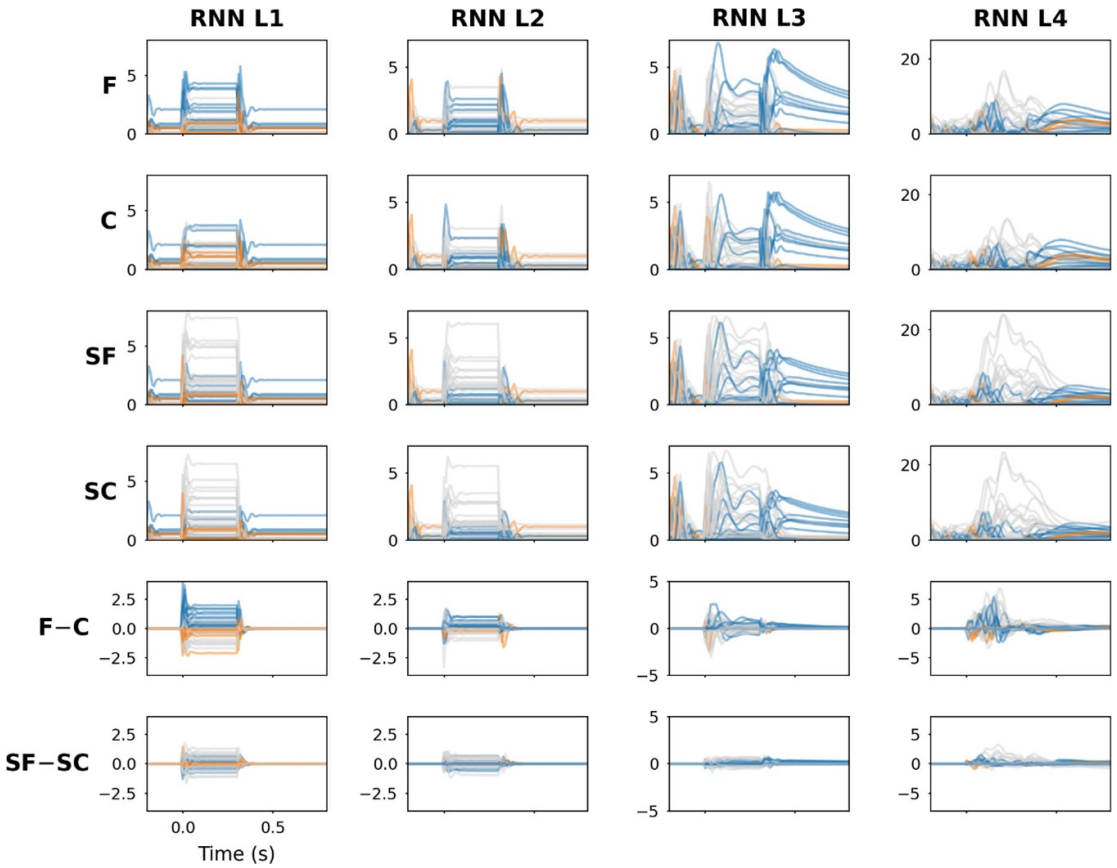

Figure 6. Hidden unit activations from the RNN section of the model. Units maximally sensitive to face stimuli are colored blue, those most sensitive to car stimuli are colored orange, and the rest are colored gray. Responses to faces (F), cars (C), scrambled faces (SF) and scrambled cars (SC) are shown, as well as waveform subtractions, F − C and SF − SC. Layer 4 activations were directly mapped onto model outputs representing ERP waveforms, so are considered to reflect candidate patterns of neural activity responsible for shaping ERP morphology.

processing of face and car images. This data is plotted in figure 6.

2.2.4. Simulated experiments

We ran simulated experiments using the final model. Synthetic images similar to those in the original EC experiment were used in these simulations. Artificial face and car images were generated with GANs and phase-scrambled to produce four image categories, each with 1000 samples. These were presented to the model in four conditions: upright, inverted, LF and HF. Image filtering was performed in the frequency domain using a filter mask containing a circle with 5 px radius smoothed with a 5 by 5 px average filter. For the low-pass filter mask, the center circle, intersecting low spatial frequencies, was equal to 1, whereas outside the center circle, intersecting higher spatial frequencies, was equal to 0. The high-pass filter mask was an inverted copy of the low-pass filter mask; i.e. 0 at the center and 1 outside the center. Model outputs were inspected and difference waveforms face minus car (F − C) and scrambled face minus scrambled car (SF − SC) were computed. The results from these experiments are plotted in figure 7. The MQ dataset was collected in a follow-up experiment to evaluate the results of simulations involving inverted images, the results of which are shown in figure 8.

2.3. Statistical analysis

Pearson's correlation coefficient (r) was used to compare ERPs from EC, MQ and CV datasets (figure 3) and RDM matrices from CNN outputs, RNN layers and both sets of ERPs (figure 4). Difference waveforms in figures 7 and 8 were evaluated by comparing the two associated sets of ERP waveforms with two-tailed Mann–Whitney U-tests at every time point, followed by Bonferroni corrections for multiple comparisons. The threshold for statistical significance was set conservatively for all statistical tests, with an alpha value of 0.001.

2.4. Software

Python 3 was used with Matplotlib 3.5.2, NeuroRA 1.1.6.8, NumPy 1.22.4, OpenCV-Python 4.6.0.66, Scikit-Learn 1.1.1, SciPy 1.8.1, and TensorFlow 2.9.1. The model and code are freely available from [https://osf.io/y36e5/?view_only=5ec917a504a1492ab0b69f4d2500f485].

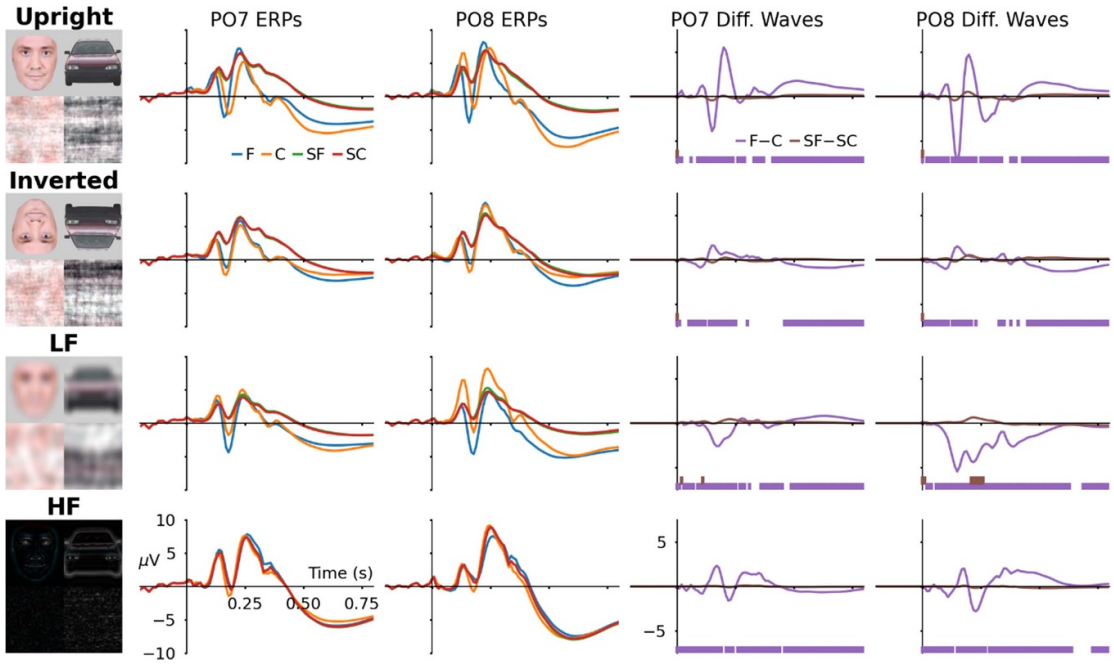

Figure 7. Simulated experiments performed with the model and synthetic images. One thousand artificial images of faces (F), cars (C), scrambled faces (SF) and scrambled cars (SC) were formatted as input arrays and fed into the model to predict corresponding ERP waveforms reflected by model outputs. Three manipulations were performed on these sets of images: they were inverted, LF and HF. Time-points with statistically significant ($p < 0.001$, with Bonferroni corrections) differences between sets of model outputs used to compute difference waveforms are illustrated with colored bars below the difference waveform plots on the right.

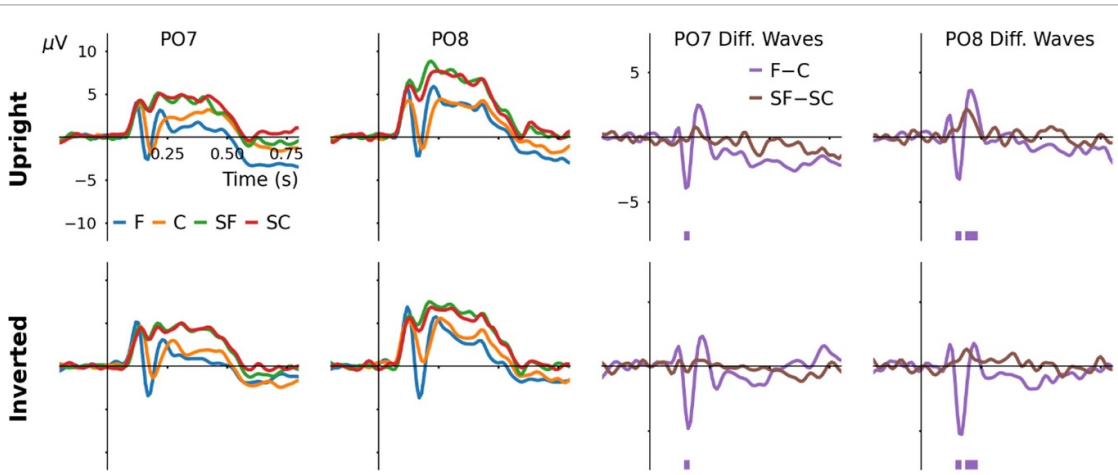

Figure 8. Validation experiment performed in a cohort of human subjects. The top row displays validation study (MQ) data from figure 3. Upright and inverted images of faces (F), cars (C), scrambled faces (SF) and scrambled cars (SC) were presented to 16 healthy adult subjects to evoke these ERP waveforms. Statistically significant results from timewise Mann–Whitney *U*-tests ($p < 0.001$, with Bonferroni corrections) comparing the two sets of ERPs used to produce difference waveforms are annotated below difference wave plots in their respective colors.

3. Results

3.1. Model outputs are strongly correlated with real ERP waveforms

Grand-average ERPs from validation study (MQ), EC and model CV datasets are plotted in the upper three rows of figure 3. Qualitatively, these three sets of ERPs are in close agreement. Correlation coefficients relating each pair of waveforms are provided in the bottom row of figure 3, separated by stimulus category. Across all stimulus responses, CV and EC waveforms were highly correlated ($r = 0.98$, $p < 0.001$), CV and MQ were slightly less correlated ($r = 0.78$, $p < 0.001$) similar to EC and MQ waveforms ($r = 81$, $p < 0.001$). It can be seen that model outputs are strongly correlated ($r > 0.85$) with real EC ERP waveforms, whereas the two sets of real ERP waveforms, and model outputs and MQ ERP waveforms, tend towards being strongly correlated ($r > 0.75$). Individual ERPs from each dataset are plotted in supplementary figures S2–S5.

3.2. Learned image vector representations are specific to known objects

CNN vector representations transformed into two principal components are visualized in figure 4. The first principal component (PC1) accounted for 45.45% and the second principal component (PC2) accounted for 24.46% of the variance between vector representations of real images. Faces and cars occupy distinct regions in this space whereas their scrambled counterparts overlap. A logistic regression classifier trained on full-size (240 element) vector representations of real images and tested on vectors representing synthetic images achieved an overall training accuracy score of 1.0 and testing accuracy score of 0. 802. Evaluating test set accuracy by image category: faces 1.0, cars 0.995, scrambled faces 0.523 and scrambled cars 0.689. In terms of misclassifications, four car image vectors were classified as scrambled faces and one was classified as scrambled car (see supplementary figure S6 for details; a random selection of synthetic face images is also shown in figure S7). Scrambled face image vectors were incorrectly classified as face 7, car 27, and scrambled car 443 times. Vectors representing scrambled cars were misclassified as face 3, car 61, and scrambled face 247 times.

3.3. Representational similarity with neural data increases along model hierarchy

Model behavior was compared with EC and MQ responses using RSA [28]. These analyses are visualized in figure 5. Comparing the lower triangle of RDM data displayed in the top row of figure 5 with those of the EC dataset produced correlation coefficients of -0.017 ($p = 0.0571$) for CNN, 0.183 for RNN layer 1, 0.293 for RNN layer 2, 0.367 for RNN layer 3, 0.485 for RNN layer 4, 0.668 for model output; where not stated, $p < 0.001$. Comparing model activations with MQ data produced correlation coefficients of 0.018 ($p = 0.04$) for CNN, 0.014 ($p = 0.11$) for RNN layer 1, 0.029 ($p = 0.001$) for RNN layer 2, 0.012 ($p = 0.18$) for RNN layer 3, -0.013 ($p = 0.13$) for RNN layer 4, 0.028 ($p = 0.0017$) for model output. The two RDMs computed from real neural data (i.e. EC vs. MQ) had $r = 0.021$ ($p = 0.0172$).

When compared over time, patterns of correlation between model RDMs from the RNN section and EC RDMs tend towards displaying twin peaks at approximately 90 ms and 200 ms post stimulus onset. In contrast, CNN outputs displayed a single peak correlation at 140 ms ($r = 0.128$, $p < 0.001$). For both overall RDMs and time-wise RDMs these correlations were proportional to layer proximity to the model output; i.e. the RDM from model output was most strongly correlated, whereas the RDM from CNN outputs were least strongly correlated with the EC RDM data. Compared with the MQ RDM data across time, CNN activity peaked in correlation at 150 ms ($r = 0.275$, $p < 0.001$), RNN layers 1–3 peaked at 180 ms ($r = 0.182$, $r = 0.147$, $r = 0.13$; all $p < 0.001$), RNN layer 4 peaked in correlation at 170 ms ($r = 0.0781$, $p < 0.001$), and model output and MQ RDMs correlated with a peak at 0.13 s ($r = 0.0545$, $p < 0.001$). Correlation between MQ and EC RDM data peaked at 180 ms ($r = 0.0676$, $p < 0.001$).

3.4. Recurrent units show differences between face and car image processing

Hidden unit activations from the RNN section are shown in figure 6. Responses to face, car, scrambled face and scrambled car stimuli are displayed in the top four rows. Face minus car and scrambled-face minus scrambled-car responses are plotted in the bottom two rows. Traces are colored according to which stimulus category elicited maximum responses from each hidden unit: face (blue), car (orange) or scrambled (gray). Substantial differences between responses to face and car stimuli are observed from outputs of layer one, the first transformation of CNN vector outputs. Differences between responses to face and car images are apparent as information propagates through the network to layer four. In comparison, differences between scrambled face and scrambled car inputs are less pronounced at every layer. Activations from layer four units are combined in a linear superposition to produce model outputs. These are therefore conceptually similar to neural sources that contribute to scalp-recorded ERP signals. The same layer four units are activated by both face and car stimuli, however, the magnitude of responses to faces tended to be greater than those of cars.

3.5. Simulations suggest patterns of differences between inverted, low-frequency and high-frequency stimuli

Results from four simulated experiments are shown in figure 7. Upright images produced an N170 response to faces that was earlier and greater amplitude than that of car images. Difference waveforms between upright face and car responses highlight this feature and its laterality towards the right hemisphere (channel PO8). Upright scrambled images did not produce significant differences in model outputs. Inverting car and face images inverted the pattern of ERP differences between their model responses. Low-frequency face stimuli elicited more negative amplitude model output relative to car stimuli, whereas high-frequency produced phasic differences in model output deflections. Inverted, low-frequency and high-frequency scrambled images did not produce substantially different outputs from the model.

3.6. Validation study is partly consistent with model simulations

ERPs obtained from the MQ validation experiment with inverted images are plotted in figure 8. The face minus car ERP difference waveform produced by responses to upright images has more well-defined

alternating polarity morphology, whereas that produced by inverted images shows a relatively pronounced negative peak and diminished positive peak. These differences reflect changes to ERP waveforms evoked by face and car images when they are inverted. In contrast, difference waveforms produced by subtracting responses to scrambled car images from those of scrambled face images do not change notably when the images are upside down. Model outputs in response to inverted scrambled images are consistent with these observations, whereas model outputs in response to inverted face and car images are not.

4. Discussion

The aim of developing this CNN-RNN model was to explore whether it can reproduce ERP responses evoked by images of faces, cars, scrambled faces and scrambled cars. This was demonstrated with CV model outputs shown in figure 1 and supplementary figures S2–S5. Maximum correlation was observed between model outputs and EC data, which is unsurprising given that samples from this dataset were used to develop the models. The validation study dataset correlated less highly with CV model outputs. This reduction is considered to reflect data variability in ERP experiments conducted in separate labs, with different participants, using slightly different equipment. However, the still relatively high correlations should be taken as an indication of the similarity between the datasets.

Vector representations learned by the CNN section distinguished between face and car stimuli but not their phase-scrambled equivalents. This is seen in two-dimensional principal component space plotted in figure 4, and also from the results of vector classification. More accurate classification of face and car image vectors compared with scrambled image vectors suggests that the model only needed to identify these three categories to reliably generate output waveforms matching ERPs. This can be explained by the fact that ERP waveforms elicited by scrambled face and scrambled car images were effectively indistinguishable. This indicates that semantic content within images is more consequential than color for producing model outputs and ERP waveforms, given that the scrambled images contained the same colors as unscrambled images in unrecognizable configurations. The research literature on N170 sensitivity to faces, other objects and non-objects is consistent with these observations [4, 6, 9, 29].

RSA [28] was used to compare neural data from EC and the validation study (MQ) with final CNN-RNN model activations in figure 5. RDMs were computed to quantify correlation between model activations at multiple levels in response to 160 different input stimuli across the four categories. When calculated over the whole epoch, correlation between model and EC RDMs increased with proximity to the model output. In contrast, model and MQ RDMs showed negligible correlation when evaluated over the whole epoch. However, when evaluated timewise over five-sample sliding windows, significant correlations between model activations and both MQ and EC RDMs emerged, principally between 150 and 180 ms, suggesting that meaningful variance in the neural responses to stimulus images is concentrated within this time period. Correlations between EC and model activations tended to be higher than some previous reports [30, 31], presumably due to the supervised learning paradigm applied to constrain our model to replicate these ERP waveforms. Although correlations between model activations and MQ data, and EC and MQ data were more closely aligned with previous results of RSA [30, 31].

Hidden units in the RNN section of the model demonstrated different patterns of activity in response to face and car stimuli, illustrated in figure 6. At the fourth hidden layer, the majority of these differences occurred before 200 ms after stimulus onset. Activations of these units are weighted together to generate model outputs and are therefore considered analogous to the activity of neural sources. We can see that the same units were activated by all stimuli, although the magnitudes of those activations were dependent upon stimulus category, with face images tending to produce larger amplitude responses. Interpreted within the context of the debate regarding face versus familiar-object specificity of N170, these findings support the view that the underlying source activity is sensitive to familiar objects rather than being specific to a particular category of objects. This should not be overstated, however, as there is inherent difficulty in separating neighboring, temporally-overlapping sources from ERP signals. Furthermore, this may reflect a tendency for the RNN's hidden units to not be stimulus-specific per-se (i.e. not solely responding to a specific category of stimuli), but instead have their amplitudes modulated by different types of stimuli.

Synthetic images generated by GANs were realistic-looking and elicited reasonable responses from the model. These were created to evaluate model performance on a large set of unseen images from the same distribution as the real images used in EC and MQ experiments. Synthetic face images produced an N170 peak with greater amplitude at channel PO8 relative to PO7. Vector representations produced by the model in response to synthetic images (semi-transparent markers in figure 4) were also clustered according to their intended category. From the simulation results illustrated in figure 7, differences between face and car images change with each type

of image manipulation, whereas differences between scrambled images remain constant. This suggests that when face and car images are sufficiently distorted the model ceases to output waveforms comparable with those in the training data. It is not unreasonable to suspect that changes to the appearance of images might similarly influence ERP waveforms [32–34].

Results from the validation experiment in figure 8 do not concur with model responses to inverted face and car images (second row of figure 7). The face N170 peak seen in real ERPs is enlarged and slightly delayed when images are inverted [5, 7, 8], whereas model outputs to inverted faces are attenuated. Comparing MQ ERPs and model outputs in response to inverted face and inverted car images (supplementary figure S8) highlights these differences. In contrast, validation experiment ERPs in response to scrambled images were relatively unaffected by inversion, which is consistent with model outputs. This is unsurprising because random phase-scrambled images do not have a 'right way up', therefore the model (and presumably the human brain) makes no distinction whether they are upright and or inverted. Considering the variable success of these predictions, we cannot assume that model outputs in response to low- and high-pass filtered images are accurate representations of ERPs that would be elicited by these images in real experiments. Nevertheless, it is likely that these image manipulations would cause some changes to ERP morphology [5, 7, 8].

In future work this model could be fine-tuned with data acquired from experiments investigating more granular aspects of human face processing, such as ambiguity [35], emotion [36], familiarity [37, 38], racial congruence [39], or wearing of partial facial coverings [40, 41], to evaluate how these affect model behavior. Doing so would provide additional analyses to assist in interpreting computational principles underlying sensitivity of brain responses to different manipulations of face stimuli. Furthermore, this modeling approach could be applied in other visually-evoked potential contexts as a model of ERP generation. Image vector representations learned by the CNN and patterns of RNN hidden unit activations that combine to produce model outputs matching ERP waveforms can be characterized, along with the generated ERPs, in efforts to gain insights into the neurophysiological computations of visual perception.

5. Conclusion

The CNN-RNN model reliably captured some of the key features of ERP responses to face, car and phase-scrambled images. These include N170 sensitivity to face images lateralized to the right hemisphere. Different latent states observed from the model became increasingly correlated with ERP responses as they were transformed from vector representations through to model output signals. This is somewhat analogous to visual information processing from the retina through to higher-order occipital cortex. However, model predictions in response to inverted face images were inconsistent with expectations and our validation experiment, highlighting an important limitation. The model could not reliably reproduce unseen patterns of neural activity in response to image manipulations. Nevertheless, further development and application of this model may be beneficial for analyzing subtler aspects of face perception and generally to explore the computational principles underlying visually evoked potentials.

Data availability statements

The event-related potential recordings and image stimuli used to develop the model are available from ERP CORE which can be accessed from https://osf.io/thsqg/. Code and data for modeling and synthetic images are available from [https://osf.io/y36e5/?view_only=5ec917a504a1492ab0b69f4d2500f485].

The data that support the findings of this study are openly available at the following URL/DOI: https://osf.io/y36e5/ and https://osf.io/thsqg/

Acknowledgments

This work was supported in part by an International Brain Research Organization (IBRO) Exchange Fellowship awarded to JAO to train with PFS. The computer and graphics processing unit used to train the models were purchased with a Grant from the Research Institute of Rangsit University (Grant Numbers 90/2561).

Author contributions

JAO: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing—original draft, Visualization, Supervision, Funding acquisition.

JW: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing—review & editing, Supervision.

AC: Validation, Investigation.

JB: Validation, Investigation.

TH: Validation, Investigation.

FA: Methodology, Software, Formal analysis.

PFS: Resources, Writing—review & editing, Supervision, Project Administration, Funding acquisition.

Conflict of interests

The authors declare no conflicts of interest.

Ethical approval

The validation study conducted at Macquarie University was reviewed and approved by the Human Research Ethics Committee.

ORCID iDs

Jamie A O'Reilly https://orcid.org/0000-0002-2250-3077
Jordan Wehrman https://orcid.org/0000-0002-9358-6092
Paul F Sowman https://orcid.org/0000-0002-3863-6675

References

[1] Leopold D A and Rhodes G 2010 A comparative view of face perception *J. Comp. Psychol.* **124** 233–51
[2] Zagury-Orly I, Kroeck M R, Soussand L and Cohen A L 2022 Face-processing performance is an independent predictor of social affect as measured by the autism diagnostic observation schedule across large-scale datasets *J. Autism Dev. Disord.* **52** 674–88
[3] Behrmann M and Avidan G 2005 Congenital prosopagnosia: face-blind from birth *Trends Cogn. Sci.* **9** 180–7
[4] Osborne K J, Kraus B, Curran T, Earls H and Mittal V A 2022 An event-related potential investigation of early visual processing deficits during face perception in youth at clinical high risk for psychosis *Schizophr. Bull.* **48** 90–99
[5] Wehrman J, Sörensen S, De Lissa P and Badcock N A 2021 EPOC outside the shield: comparing the performance of a consumer-grade eeg device in shielded and unshielded environments *Biomed. Phys. Eng. Express* **7** 025010
[6] Rossion B and Caharel S 2011 ERP evidence for the speed of face categorization in the human brain: disentangling the contribution of low-level visual cues from face Perception *Vis. Res.* **51** 1297–311
[7] Bossi F, Premoli I, Pizzamiglio S, Balaban S, Ricciardelli P and Rivolta D 2020 Theta- and gamma-band activity discriminates face, body and object perception *Front. Hum. Neurosci.* **14** 74
[8] Bentin S, Allison T, Puce A, Perez E and McCarthy G 1996 Electrophysiological studies of face perception in humans *J. Cogn. Neurosci.* **8** 551–65
[9] Torriero S, Mattavelli G, Gerfo E L, Lauro L R, Actis-Grosso R and Ricciardelli P 2019 FEF excitability in attentional bias: a TMS-EEG study *Front. Behav. Neurosci.* **12** 333
[10] Ricciardelli P, Ro T and Driver J 2002 A left visual field advantage in perception of gaze direction *Neuropsychologia* **40** 769–77
[11] Richards B A *et al* 2019 A deep learning framework for neuroscience *Nat. Neurosci.* **22** 1761–70
[12] Richards B, Tsao D and Zador A 2022 The application of artificial intelligence to biology and neuroscience *Cell* **185** 2640–3
[13] Lindsay G W 2021 Convolutional neural networks as a model of the visual system: past, present, and future *J. Cogn. Neurosci.* **33** 2017–31
[14] O'Reilly J A 2022 Recurrent neural network model of human event-related potentials in response to intensity oddball stimulation *Neuroscience* **504** 63–74
[15] O'Reilly J A, Angsuwatanakul T and Wehrman J 2022 Decoding violated sensory expectations from the auditory cortex of anaesthetised mice: hierarchical recurrent neural network depicts separate 'danger' and 'safety' units *Eur. J. Neurosci.* **56** 4154–75
[16] O'Reilly J A, Wehrman J and Sowman P F 2022 A guided tutorial on modelling human event-related potentials with recurrent neural networks *Sensors* **22** 9243
[17] O'Reilly J A 2022 Modelling mouse auditory response dynamics along a continuum of consciousness using a deep recurrent neural network *J. Neural Eng.* **19** 056023
[18] Shi J, Wen H, Zhang Y, Han K and Liu Z 2018 Deep recurrent neural network reveals a hierarchy of process memory during dynamic natural vision *Hum. Brain Mapp.* **39** 2269
[19] Xu G, Ren T, Chen Y and Che W 2020 A one-dimensional CNN-LSTM model for epileptic seizure recognition using EEG signal analysis *Front. Neurosci.* **14** 1253
[20] Grill-Spector K, Weiner K S, Gomez J, Stigliani A and Natu V S 2018 The functional neuroanatomy of face perception: from brain measurements to deep neural networks *Interface Focus* **8** 20180013
[21] Kappenman E S, Farrens J L, Zhang W, Stewart A X and Luck S J 2021 ERP CORE: an open resource for human event-related potential Research *Neuroimage* **225** 117465
[22] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets *Advances in Neural Inf. Proc. Systems* p 27
[23] Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J and Aila T 2020 Training generative adversarial networks with limited data *Advances in Neural Information Processing Systems* (*December 2020*) (https://doi.org/10.48550/arxiv.2006.06676)
[24] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T 2020 Analyzing and improving the image quality of stylegan *Proc. IEEE Comput. Society Conf. computer vision and pattern Recognition* vol 2019 pp 8107–16
[25] Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S 2017 GANs trained by a two time-scale update rule converge to a local nash equilibrium *Advances in Neural Information Processing Systems* (*Decemeber 2017*) pp 6627–38
[26] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift *Proc. 32nd Int. Conf. on Machine Learning, ICML 2015; Int. Machine Learning Society (IMLS)* vol 1 pp 448–56
[27] Melinsca M, Prentasic P and Loncaric S 2015 Retinal vessel segmentation using deep neural networks *Proc. VISAPP 2015–10th Int. Conf. on Computer Vision Theory and Applications; VISIGRAPP, Proc.* (*1 January 2015*) vol 1 pp 577–82
[28] Kriegeskorte N, Mur M and Bandettini P 2008 Representational similarity analysis—connecting the branches of systems neuroscience *Front. Syst. Neurosci.* **2** 4
[29] Rousselet G A, Pernet C R, Bennett P J and Sekuler A B 2008 Parametric study of EEG sensitivity to phase noise during face processing *BMC Neurosci.* **9** 1–22
[30] Kiat J E, Hayes T R, Henderson J M and Luck S J 2022 Rapid extraction of the spatial distribution of physical saliency and semantic informativeness from natural scenes in the human brain *J. Neurosci.* **42** 97–108
[31] He T, Boudewyn M A, Kiat J E, Sagae K and Luck S J 2022 Neural correlates of word representation vectors in natural language processing models: evidence from representational similarity analysis of event-related brain potentials *Psychophysiology* **59** e13976
[32] Male A G, O'Shea R P, Schröger E, Müller D, Roeber U and Widmann A 2020 The quest for the genuine visual mismatch negativity (VMMN): event-related potential indications of deviance detection for low-level visual features *Psychophysiology* **57** e13576
[33] Johannes S, Münte T F, Heinze H J and Mangun G R 1995 Luminance and spatial attention effects on early visual processing *Cogn. Brain Res.* **2** 189–205
[34] Lacroix A, Harquel S, Mermillod M, Vercueil L, Alleysson D, Dutheil F, Kovarski K and Gomot M 2022 The predictive role of low spatial frequencies in automatic face processing: a

visual mismatch negativity investigation *Front. Hum. Neurosci.* 16 101
[35] Abubshait A, Momen A and Wiese E 2020 Pre-exposure to ambiguous faces modulates top-down control of attentional orienting to counterpredictive gaze cues *Front. Psychol.* **11** 2234
[36] Petrucci M and Pecchinenda A 2017 The role of cognitive control mechanisms in selective attention towards emotional stimuli *Cogn. Emot.* **31** 1480–92
[37] Chauhan V, Kotlewska I, Tang S and Gobbini M I 2020 How familiarity warps representation in the face space *J. Vis.* **20** 18
[38] Wiese H, Hobden G, Siilbek E, Martignac V, Flack T R, Ritchie K L, Young A W and Burton A M 2022 Familiarity is familiarity is familiarity: event-related brain potentials reveal qualitatively similar representations of personally familiar and famous faces *J. Exp. Psychol.: Learn. Mem. Cogn.* **48** 1144–64
[39] Sessa P and Dalmaso M 2016 Race perception and gaze direction differently impair visual working memory for faces: an event-related potential study *Soc. Neurosci.* **11** 97–107
[40] Noyes E, Davis J P, Petrov N, Gray K L H and Ritchie K L 2021 The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers *R. Soc. Open Sci.* **8** 201169
[41] Calbi M, Langiulli N, Ferroni F, Montalti M, Kolesnikov A, Gallese V and Umiltà M A 2021 The consequences of COVID-19 on social interactions: an online study on face covering *Sci. Rep.* **11** 1–10