Received: 25 February 2020 | Revised: 19 May 2020 | Accepted: 29 June 2020

DOI: 10.1002/hbm.25136

RESEARCH ARTICLE

WILEY

Long short-term memory-based neural decoding of object categories evoked by natural images

Wei Huang[1] | Hongmei Yan[1] | Chong Wang[1] | Jiyi Li[1] | Xiaoqing Yang[1] | Liang Li[1] | Zhentao Zuo[2] | Jiang Zhang[3] | Huafu Chen[1]

[1]The MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, People's Republic of China

[2]State Key Laboratory of Brain and Cognitive Science, Beijing MR Center for Brain Research, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China

[3]Department of Medical Information Engineering, Sichuan University, Chengdu, China

Correspondence

Hongmei Yan and Huafu Chen, MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China.
Email: hmyan@uestc.edu.cn (H. Y.) and chenhf@uestc.edu.cn (H. C.)

Zhentao Zuo, State Key Laboratory of Brain and Cognitive Science, Beijing MR Center for Brain Research, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China.
Email: ztzuo@bcslab.ibp.ac.cn

Funding information

Chinese Academy of Sciences Strategic Priority Research Program B, Grant/Award Number: XDB32010300; National Major Scientific Instruments and Equipment Development Project, Grant/Award Number: ZDYZ2015-2; Ministry of Science and Technology, Grant/Award Number: 2015CB351701; National Natural Science Foundation of China, Grant/Award Numbers: 61876114, 31671133, 31730039, U1808204, 61533006, 61573080, 61773094

Abstract

Visual perceptual decoding is one of the important and challenging topics in cognitive neuroscience. Building a mapping model between visual response signals and visual contents is the key point of decoding. Most previous studies used peak response signals to decode object categories. However, brain activities measured by functional magnetic resonance imaging are a dynamic process with time dependence, so peak signals cannot fully represent the whole process, which may affect the performance of decoding. Here, we propose a decoding model based on long short-term memory (LSTM) network to decode five object categories from multitime response signals evoked by natural images. Experimental results show that the average decoding accuracy using the multitime (2–6 s) response signals is 0.540 from the five subjects, which is significantly higher than that using the peak ones (6 s; accuracy: 0.492; $p < .05$). In addition, from the perspective of different durations, methods and visual areas, the decoding performances of the five object categories are deeply and comprehensively explored. The analysis of different durations and decoding methods reveals that the LSTM-based decoding model with sequence simulation ability can fit the time dependence of the multitime visual response signals to achieve higher decoding performance. The comparative analysis of different visual areas demonstrates that the higher visual cortex (VC) contains more semantic category information needed for visual perceptual decoding than lower VC.

KEYWORDS

deep learning, fMRI, LSTM, visual cortex, visual perceptual decoding

1 | INTRODUCTION

Exploring how the brain represents visual information from our daily visual experiences is such a meaningful and challenging task that it has attracted many neuroscientists. To date, the ways our brain responds and encodes the visual world have been widely studied, and many neural mechanisms have been elucidated. With the rapid development of some good noninvasive technologies that measure brain activity with reasonable spatial and temporal resolution, such as functional magnetic resonance imaging (fMRI) and electroencephalogram (EEG), researchers have tried to figure out whether brain activity can be used to identify what people see or perceive. This process is called visual perceptual decoding or "brain-reading" (Cox & Savoy, 2003; Gallagher et al., 2000; Güçlütürk et al., 2017; Norman, Polyn, Detre, & Haxby, 2006; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014).

Recent studies have developed lots of methods to classify or identify visual stimuli from evoked fMRI/EEG responses. These category-based visual perceptual decoding researches explored the perceived categories of natural images (Behroozi & Daliri, 2014; Behroozi, Daliri, & Shekarchi, 2015; Carlson, Schrater, & He, 2003; Cox & Savoy, 2003; Haxby et al., 2001; Jafakesh, Jahromy, & Daliri, 2016; Jahromy & Daliri, 2017; Kamitani & Tong, 2005; Kay, Naselaris, Prenger, & Gallant, 2008; Song, Zhan, Long, Zhang, & Yao, 2011; Tafreshi, Daliri, & Ghodousi, 2019; Taghizadeh-Sarabi, Daliri, & Niksirat, 2015; Torabi, Zareayan Jahromy, & Daliri, 2017), real inner thoughts (Yang et al., 2014), imagined categories (Naselaris, Olman, Stansbury, Ugurbil, & Gallant, 2015) and categories occurring in dreams (Horikawa, Tamaki, Miyawaki, & Kamitani, 2013), and so on. These studies strove to improve the decoding performance of their research fields from different experimental protocols and methods. So far, from the aspect of fMRI, since the research of category-based decoding is affected by many factors, such as the low signal-to-noise ratio of fMRI signals and limited sample size due to the constraints of fMRI data collection, better data processing methods, and decoding models are demanded for further explorations.

Most traditional category-based visual decoding methods attempt to establish a mapping relationship between the peak response of fMRI signals and the object category (Carlson et al., 2003; Haxby et al., 2001; Haynes & Rees, 2005; Kamitani & Tong, 2005). However, the visual response signals evoked by visual stimuli have a delay (Cox & Kable, 2014; Zong, Kim, & Kim, 2012) from human brain activity measured by fMRI. Because fMRI indirectly reflects the neural activities of the cerebral cortex by measuring changes in blood oxygen concentration (Lippert, Steudel, Ohl, Logothetis, & Kayser, 2010), when viewing natural images, the fMRI activities of the cerebral cortex are distributed over a period of time. Temporal decoding methods for EEG and MEG have offered the potential to utilize the temporal dynamics in object decoding (Barragan-Jason, Cauchoix, & Barbeau, 2015; Carlson, Hogendoorn, Kanai, Mesik, & Turret, 2011). However, few the fMRI-based decoding studies use the time dependence of response patterns to improve the decoding performance (Contini, Wardle, & Carlson, 2017). The main reason may be that, when the visual stimulus presents, due to the response delay, hemodynamic response arises first and then falls down gradually. The peak signal of hemodynamic response function (HRF) shows the strongest response to the stimuli, and the signals before and after the peak contain weaker response and more noise than the peak ones. In our opinion, although the fMRI signals before and after the peak are not as strong as the peak ones, they still contain lots of stimuli-related neural activities which can be used for decoding although the accuracy of prepeak or postpeak signals may be lower than that of peak ones. Our result shown also proved that prepeak signals can also be used for decoding to a certain degree. Here, we reasonably deduce that the dynamic process of fMRI may help improve decoding performance if appropriate methods are adopted to make full use of the time-dependence signals. As we know, long short-term memory network (LSTM) (Hochreiter & Schmidhuber, 1997) is a good approach to handle dynamic changes of the time-dependence signals. Therefore, we proposed a LSTM-based decoding model to mine the multitime visual response signals evoked by the natural images, so as to achieve the better decoding of object category. We managed to apply the LSTM-based decoding model to classify the multitime visual response signals evoked by five types of natural images (horses, buildings, flowers, fruits, and landscapes). The experimental results show that the decoding accuracy using multitime visual response signals is significantly higher than that using single-time or shorter time visual response signals.

In addition, the decoding performances of the five object categories by different methods are fully compared. These decoding methods are divided into neural nets and traditional models. Neural nets include LSTM, recurrent neural network (RNN), and gated recurrent unit (GRU). Traditional models include Naive Bayes (NB), k-nearest neighbor (KNN), adaboost (Ada), random forest (RF), and support vector machine (SVM). The experimental results show that the multitime visual response signals evoked by the natural images contain more semantic category information, and the LSTM-based and GRU-based decoding model can efficiently utilize the time dependence of fMRI, yielding better decoding performance. Finally, to investigate the function of different visual areas in brain decoding, we also compared the decoding accuracies of different visual areas signals. The results show that the higher visual cortex (VC) shows higher decoding accuracy than the lower VC. Consistent with the current understanding of the visual processing hierarchy, higher-level visual features dominate decodable semantic categories (Horikawa & Kamitani, 2017).

2 | MATERIALS

2.1 | Subjects

Five healthy volunteers (three males and two females, 23–27 years) took part in the experiment. All subjects were neurologically healthy, right-handed, and had normal or corrected-to-normal vision. All subjects provided written informed consent before the experiments, and protocols were approved by the Institutional Review Board of the Institute of Biophysics, Chinese Academy of Sciences. In the experiment, all subjects were asked to keep their heads and body still and look at the center of the screen.

2.2 | Visual stimulus and experimental design

The 2,750 natural images were taken from ImageNet (Deng et al., 2009). These natural images are from five categories, namely, "horse," "building," "flower," "fruit," and "landscape," with 550 images in each category. Each image is a color image with a resolution of 256 × 256. The visual stimulus program was created with the e-prime programming software. Visual stimuli were rear-projected onto a screen placed in the scanner cavity using an LCD projector. Two types of experimental sessions were performed to measure the response signals: (a) the bar retinotopic mapping session in Figure 1a, and (b) the natural image session in Figure 1b.

The bar retinotopic mapping session was used to delineate the borders between visual cortical areas. The bars are made up of a checkerboard pattern (spatial length: 20°; spatial width: 20°; temporal frequency: 10 Hz) with 100% contrast. Four bar orientations and two different motion directions were generated for a total of eight different bar configurations within a given scan (Dumoulin & Wandell, 2008). Each bar configuration contained 22 equidistant spatial positions. The bar retinotopic mapping session included four repeated runs, and each run had 176 trials. Each stimulus trial flashed 2 s followed by the next stimulus trial. Extra rest periods were added at the beginning (12 s) and the end (12 s) of each run. Each bar run lasted (12 s + 176 trials × 2 s + 12 s = 6 min 16 s).

In the natural image session, each run had 50 stimulus trials. For each trial, one natural image flickered for 2 s (spatial length: 20°; spatial width: 20°; temporal frequency: 5 Hz) followed by a random 4–8 s (probability random integer; mean, 6 s) intervening rest period. Extra rest periods were added at the beginning (12 s) and the end (8 s) of each run. Each run lasted (12 s + 50 trials × [2 s + 6 s] + 8 s = 7 min). In each stimulus trial, a natural image was presented on a gray background with a white fixation cross (0.8° × 0.8°). Fifty-five runs were executed, and a total of 2,750 nature images were presented to each subject. Each run randomly selected 10 images from every category of the natural image dataset (flowers, fruits, horses, buildings, and landscapes). The 50 images of each run were presented to subjects in pseudorandom order. During the experiment, the subjects were instructed to fixate the small white cross at the center of the images without moving their bodies. In addition, they were asked to focus attention and try to understand the images. Each subject performed a total of 59 runs (4 bar retinotopic runs and 55 natural image runs) in the experiment. The scanning went for 8 days in 2 weeks for each subject. The 4 bar retinotopic runs finished in 1 day, and the 55 runs finished in 7 days, among them, 8 runs for each day of the first 6 days, and 7 runs for the last day. The 59 runs and fMRI data in the experiment are available on http://www.neuro.uestc.edu.cn/vccl/data/Huang2020_Article_Perception-to-ImageReconstruct.html.

2.3 | MRI scanning parameters and data preprocessing

MRI data were acquired with a 3-T Prisma$^{\text{fit}}$ scanner (Siemens, Erlangen, Germany) at the Institute of Biophysics, Chinese Academy of Sciences using a 20-channel head–neck coil. An interleaved multiband T2*-weighted gradient-echo echo-planar imaging (Auerbach, Xu, Yacoub, Moeller, & Uğurbil, 2013; Moeller et al., 2010) scan was performed to acquire functional images to cover the entire occipital lobe (TR, 1,000 ms; TE, 31.2 ms; nominal flip angle, 50°; field of view [FOV], 194 × 194 mm^2; voxel size, 1.8 × 1.8 × 1.8 mm^3; slice gap, 0 mm; multiband, 4; number of slices, 48). T1-weighted magnetization-prepared rapid-acquisition gradient-echo fine-structural images of the whole-head were also acquired (TR, 2,300 ms; TE, 3.49 ms; TI, 1,050 ms; flip angle, 8°; FOV, 256 × 256 mm; voxel size, 1.0 × 1.0 × 1.0 mm^3). The cortical surface was reconstructed at the white/gray matter border and rendered as a smoothed 3D surface. The first 12-s scans of each run were discarded to avoid MRI scanner instability. The acquired fMRI data underwent slice-timing correction and three-dimensional head motion correction by SPM8 (Penny, Friston, Ashburner, Kiebel, & Nichols, 2011). These data were then coregistered to the sliced high-resolution anatomical images and then reinterpolated to 3 × 3 × 3-mm voxels.

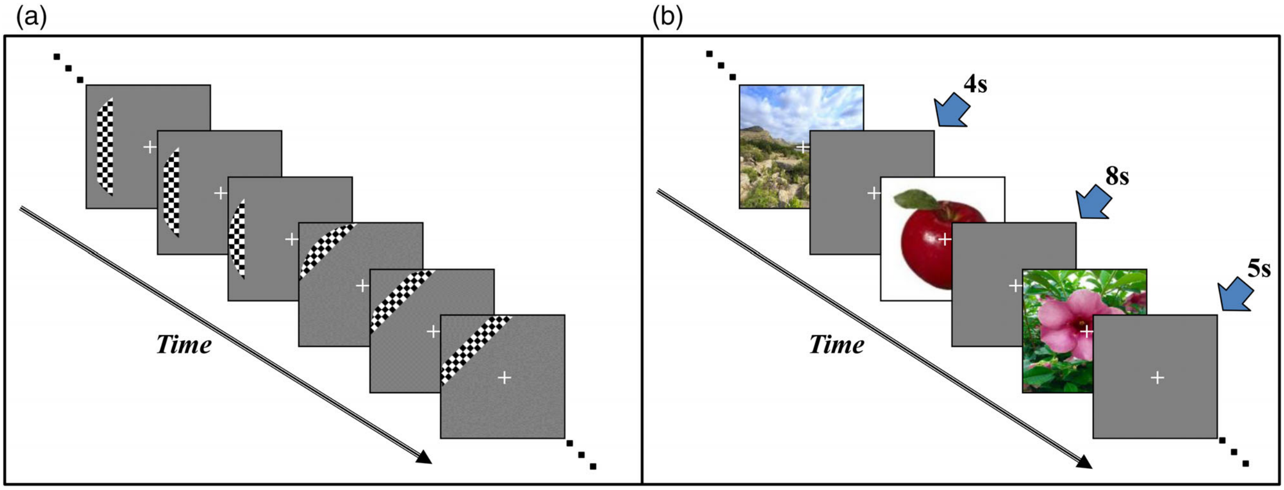


FIGURE 1 Sequence of visual stimuli. (a) The bar retinotopic mapping session. (b) The natural image session

3 | METHODOLOGY

3.1 | Population receptive field model

In the bar retinotopic mapping session, the method of estimating the neuronal population receptive field (pRF) in the human VC was based on a previous study (Dumoulin & Wandell, 2008). For each voxel, a two-dimensional simple gaussian pRF, $g(x, y)$, is defined by three parameters, x_0, y_0, and σ,

$$g(x,y) = e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}}$$

where (x_0, y_0) is the center and σ is the Gaussian spread (*SD*). For visual stimulus $s(x, y, t)$, the pRF response r_t for a single voxel is calculated by the following formula:

$$r(t) = \sum_{x,y} s(x,y,t)g(x,y)$$

Then, the predicted BOLD signal $p(t)$ is obtained by convolving $r(t)$ with a model of the HRF ($h(t)$) using the following formula,

$$p(t) = r(t)\text{⍰}h(t)$$

Assuming a linear relationship between the predicted BOLD signal and the fMRI signal $y(t)$ can be described as:

$$y(t) = p(t)\beta + \varepsilon$$

where β is a scaling factor and ε is a noise. The goodness-of-fit is estimated by computing the residual sum of squares (RSS) between the prediction $p(t)$ and the data $y(t)$. We calculated this error term allowing for a scale factor β that accounts for the unknown units of the fMRI signal,

$$\text{RSS} = \sum_t (y(t) - p(t)\beta)^2$$

The optimal pRF parameters were found by minimizing the RSS using a coarse-to-fine search. A previous study contained more details (Dumoulin & Wandell, 2008).

3.2 | Definition of visual areas

After preprocessing the fMRI data from the bar retinotopic experiment, SamSrf (SamPenDu, 2017) was used to perform retinotopic analysis. Occipital lobe is defined by SamSrf as the entire VC. Visual cortical boundaries were depicted on the spherical surface to visualize and label the visual region of interest in Figure 2. The eccentricity and the polar maps were mapped to the spherical surface. V1 contained a full hemifield map and was located fairly accurately within the calcarine sulcus (Brewer, Liu, Wade, & Wandell, 2005; Fishman, 1997). Thus, it spanned from a green stripe on the cuneus (dorsal bank of calcarine) through the blue in the depth of the calcarine to a red stripe on the lingual gyrus (ventral bank). V2 and V3 were both segregated into two quadrant field maps, one in the ventral cortex and the other in the dorsal cortex (Burkhalter, Felleman, Newsome, & van Essen, 1986). Therefore, V2d ran from the green stripe of the V1 border to the middle of the blue stripe. V3d then ran from the blue stripe to the next green stripe. Conversely, V2v ran from the red stripe of the V1 border to the blue stripe, and V3v ran from the blue strip to the next red stripe (Dougherty et al., 2003; Hubel & Wiesel, 1965). The divided V1, V2, and V3 are projected to volume space in Figure 3. Then, V1, V2, and V3 are combined as the lower VC (LVC). The remaining area of VC removing LVC is defined as the higher VC (HVC). The number of voxels in each VC is displayed in Supplementary Table S1.

3.3 | LSTM network

Neural networks in fully connected form do not handle the dynamic process very well. To enable the neural networks in fully connection to process sequence data at multiple time points, Waibel et al. proposed the time-delay neural network model (Werbos, 1990). Subsequently, an RNN was developed, and a back-propagation through time algorithm was derived theoretically to train the model. When using an RNN to process sequence data, the gradient of the loss function needs to be back-propagated through time. A large number of researchers have shown that when the sequence is very long, there will be problems with gradient disappearance or gradient explosion (Hochreiter & Schmidhuber, 1997). LSTM networks have been developed to solve this problem. Compared with the RNN, the key point of the LSTM model is to introduce a gating mechanism to control the flow of information. At time t, three gates are introduced for the input information, the memory information, and the output information, namely, the input gate (i_t), the forgot gate (f_t), and the output gate (o_t), respectively. The elements of the three gates are numbers between [0, 1]. At time t, LSTM is updated as follows:

$$\left(\begin{array}{c} i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1} + b_i) \\ f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f) \\ o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_{t-1} + b_o) \\ c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1}) \\ h_t = o_t \odot \tanh(c_t) \end{array}\right)$$

The input gate (i_t), the forgot gate (f_t), and the output gate (o_t) are determined by three factors, input (x_t), hidden status (h_{t-1}), and cell status (c_{t-1}), where σ denotes the logistic sigmoid function. Then, the hidden state and the cell state at the current moment are updated according to the three gates (i_t, f_t, o_t) at the current moment, the hidden state (h_{t-1}) and the cell state (c_{t-1}) at the previous moment. In short, the input gate controls how much new information is added; the forget gate controls the extent to which the previous state is forgotten; and the output gate controls how much of the current state is

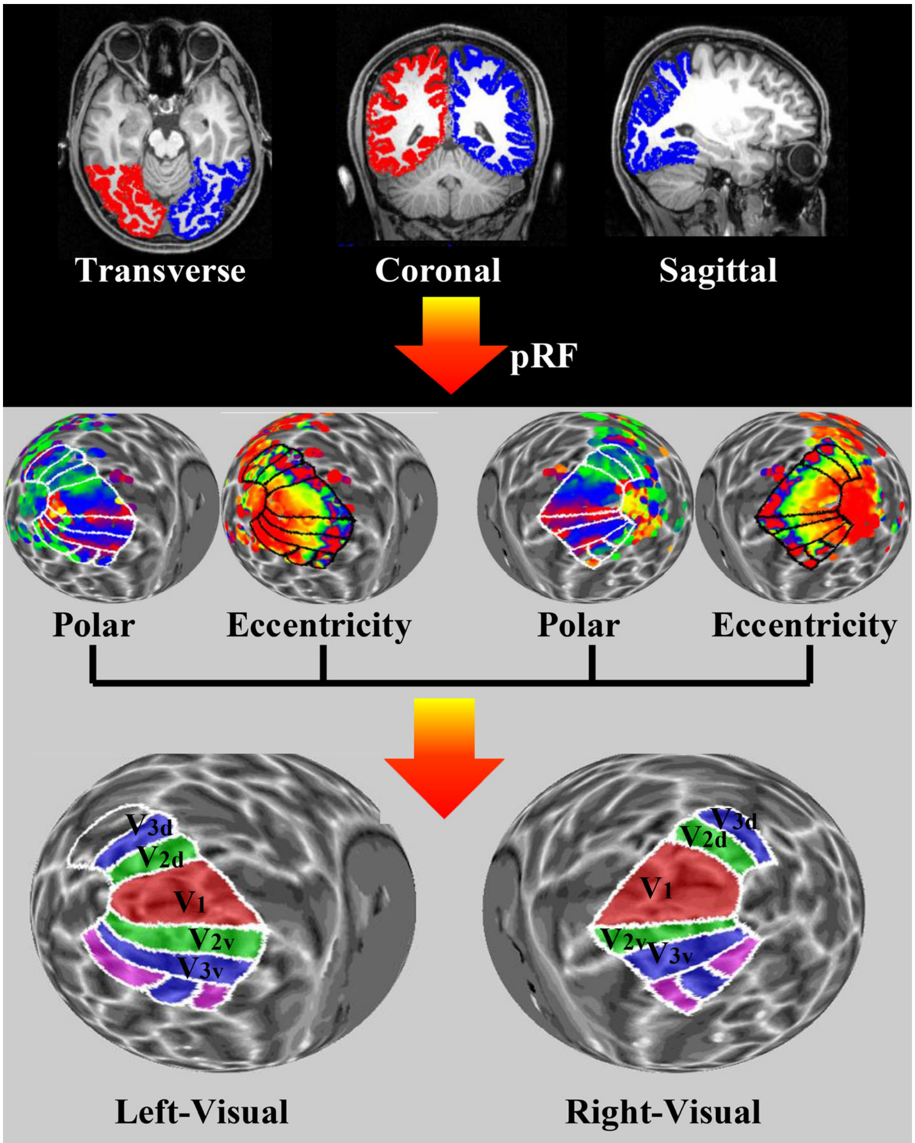


FIGURE 2 Process of dividing visual areas

filtered out. Compared with feed-forward neural networks, the LSTM model can better simulate the temporal series problem. Compared to RNN, the LSTM model alleviates the problem of gradient disappearance and gradient explosion to some extent. fMRI is an indirect measurement of neural activity through the changes in blood flow. Because blood flow change follows the HRM, which is time dependent, so the fMRI signal is also time dependent. The fMRI peak signal is delayed approximately 4–6 s relative to the stimulus (Miyawaki et al., 2008; Szaflarski et al., 2010). In this paper, based on the capability of LSTM in processing dynamic changes of the time-dependence signals, the LSTM model is used to decode the response signals at multiple time points when viewing natural images.

3.4 | The LSTM-based decoding model

From the 55 natural image sessions, five verification sessions and five test sessions were randomly selected. The remaining 45 sessions were used as the training session. Each session contained 50 natural images and corresponding fMRI response signals in the VC. The total sample set was divided into a training subset, a verification subset, and a test subset. The sizes of the three subsets were 2,250; 250; and 250, respectively. There are approximately 30,000 voxels throughout the occipital cortex, but not all of them encode visual image stimuli. Therefore, a rough voxel selection within the VC was performed first. An F-score feature selection algorithm was used to calculate the *F*-value of each voxel (Chen & Lin, 2006; Huang et al., 2018; Polat & Güneş, 2009). The higher the *F*-value of the voxel, the better the ability to discriminate visual perceptual category. Five VCs are defined: V1, V2, V3, LVC, HVC, and VC. The number of voxels is different for different subjects and different visual areas. In order to facilitate subsequent calculations, the number of voxels in each VC from different subjects is unified to the same number 2,000. The VC (V1, V2, or V3) which contains less than 2,000 voxels was upsampled by nearest neighbors to 2,000 voxels. The VC (LVC, HVC, or VC) with more than 2,000 voxels was performed by the F-score feature selection

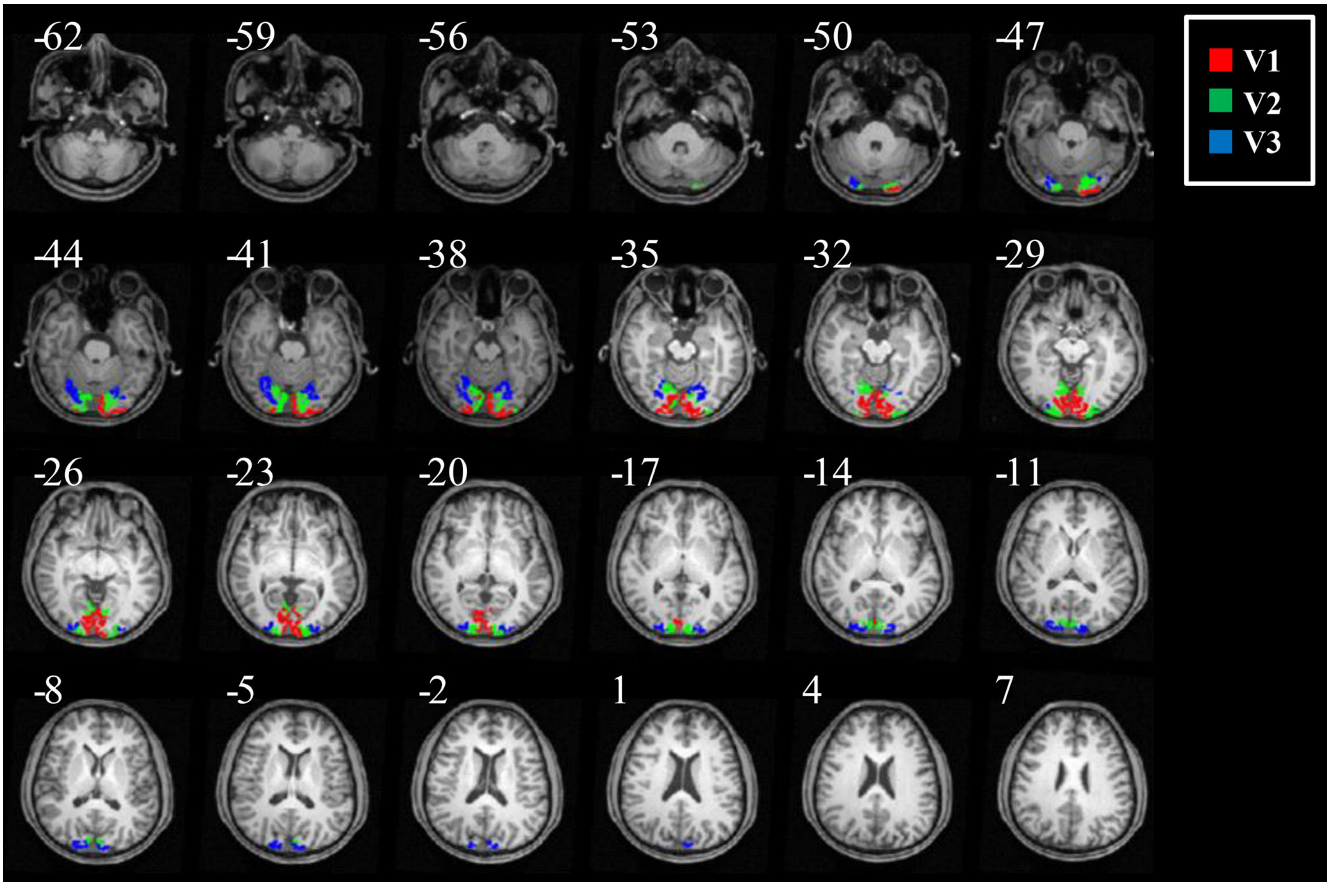
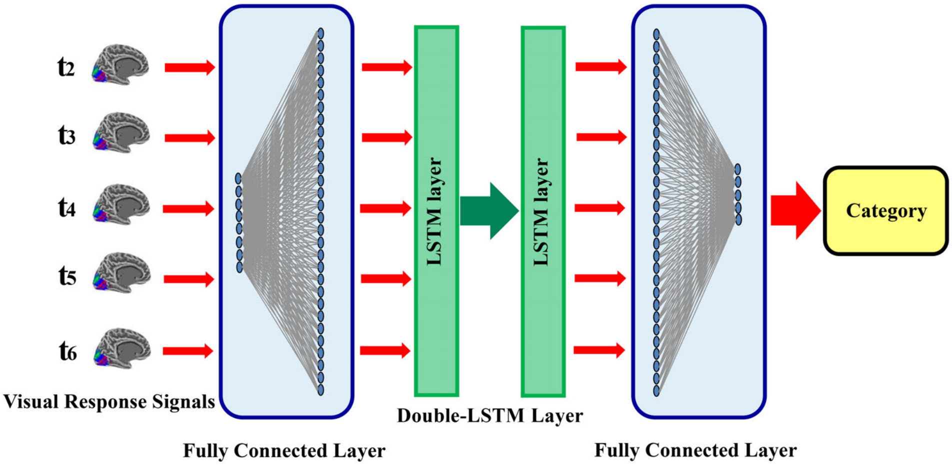


FIGURE 3 A typical volunteer's lower visual cortex (V1, V2, and V3) spatial location in volume space

FIGURE 4 The long short-term memory (LSTM)-based decoding model



algorithm and the first 2,000 voxels were chosen. Here, the number 2,000 is determined by the precalculation performance from Subject 1 with VC signals. We tested the decoding performance (Subject 1) with different numbers of voxels by LSTM. The precalculation results show that the decoding accuracy increases as the number of voxels increases from 100 to 2,000. However, the accuracy does not improve obviously when the number of voxels increases from 2,000 to 5,000. Therefore, we chose 2,000 as the number of voxels. Note that feature selection is performed only on the training subset.

The 2–6 s visual response signals measured by fMRI after the appearance of the natural image are selected as input to the LSTM-based decoding model. The LSTM-based decoding model includes two fully connected layers and a double-LSTM-layer module. Each layer consists of 5,000 units. The LSTM-based decoding model is illustrated in Figure 4. First, the LSTM-based decoding model maps the 2,000–5,000 units of the first fully connected layer. Then, the output of the first fully connected layer is processed by the double-LSTM-layer and the last fully connected layer. Finally, the output of the

LSTM-based decoding model is five category units. The softmax nonlinear mapping function is applied to the 5 units in the output layer to obtain the probability distribution of visual perceptual category. The cross entropy between the prediction label and the real label of the natural image is calculated, and the Adam (Kingma & Ba, 2014) optimization algorithm is used to optimize the LSTM-based decoding model. We determined the optimal hyperparameters based on the performance of the verification set. In brief, the crucial hyperparameters of the LSTM-based decoding model include the number of units in the first fully connected layer (5,000), the number of units in the last fully connected layer (5), the number of hidden units within double-LSTM layer (5,000), the parameter settings of Adam optimizer ($\beta_1 = .9$, $\beta_2 = .999$, $\varepsilon = 10^{-8}$), and the learning rate (0.0001).

4 | RESULTS

4.1 | Performance of the LSTM-based decoding model

Five categories of natural images, including horse, building, flower, fruit, and landscape, are selected as stimuli to evoke the functional signals of VC. Natural image categories are decoded using multitime visual response signals measured by fMRI. The LSTM-based decoding model receives the multitime visual response signals and simulates the time dependence of these signals. To demonstrate that natural images can be decoded progressively using an LSTM-based decoding model, we first show the training processes from the five subjects in Figure 5a–e. This result shows that the decoding accuracy of test and validation data first increases and then progressively moves into a steady state in the late iterative process. When the decoding accuracy reaches a steady state, the accuracies of test data from the five subjects are approximately 0.60, 0.60, 0.40, 0.60, and 0.46 (the chance level is 0.2), respectively. In addition, the loss curves of training, test, and validation data from the five subjects are shown in Figure 5f–j. From the accuracy curves, it can be intuitively seen that the LSTM-based decoding model can effectively decode the five object categories.

4.2 | Comparison of decoding performance with different durations

To clearly demonstrate that the LSTM-based decoding model can make use of the time dependence to improve the decoding performance, the decoding accuracy with different durations is compared in Figure 6. In order to make the visual response pattern of current image unaffected by the next image, the maximum of time window is set as 6 s (image 2 s + min rest 4 s). The “2–3 s,” “2–4 s,” “2–5 s,” and “2–6 s,” respectively, indicate the visual response signals of the second–third, second–fourth, second–fifth, and second–sixth seconds from human brain activity after the appearance of the natural images in the experiment. The yellow, red, green, and blue curves, respectively, represent the decoding accuracy obtained by using the visual response signals of the second–third, second–fourth, second–fifth, and second–sixth seconds as the input for the LSTM-based decoding model. The decoding accuracy curve patterns from the five subjects are similar across different durations. The performance obtained using the visual response signals of 2–6 s is better than that using signals of 2–3, 2–4, and 2–5 s. The visual response signal of 2–3 s shows the lowest decoding accuracy. In addition, we also compared the decoding accuracies obtained from the visual response signals of 5–6, 4–6, 3–6, and 2–6 s, shown in Figure 7. Similarly, the performance of 2–6 s is better than that of 5–6, 4–6, and 3–6 s. However, the difference between the decoding accuracy of 5–6, 4–6, and 3–6, and 2–6 s is smaller than that of 2–3, 2–4, and 2–5, and 2–6 s. The result shows that 2–6 s response signals provide the most semantic category information for decoding the five object categories. Besides, the visual response signals at the sixth second and close to the sixth second show greater contributions to decoding.

Finally, based on the LSTM decoding model, we compared the decoding performance of the single-time visual response signals (2, 3, 4, 5 or 6 s) and the multitime response signals (2–3, 2–4, 2–5, 5–6, 4–6, 3–6, or 2–6 s). The results are shown in Supplementary Table S2. The results show that the decoding accuracy of the multitime response signals is higher than that of the corresponding single-time signals, which also proves the positive effect of multitime signals in decoding.

4.3 | Comparison of decoding performance with different methods

To demonstrate the advantages of the proposed LSTM-based decoding model in the decoding tasks, five traditional models (NB, KNN, Ada, RF, and SVM) are used for the comparison. Note that, the multitime visual response signals are flattened and jointed together as input for these models. The five traditional models are adopted directly from the scikit-learn (Pedregosa et al., 2011) package. In addition, two neural networks (RNN-based and GRU-based) are also used for the comparison, and their training method is exactly the same as the LSTM-based decoding model.

The decoding accuracies of the different methods are calculated using 2–6 s multitime visual response signals from the VC. In order to compare different methods more reasonably, different parameters of the traditional methods are adjusted to obtain the best accuracy. The best decoding accuracies obtained by the eight methods are shown in Figure 8. The accuracies of the five subjects using the LSTM-based decoding model are 0.612, 0.608, 0.400, 0.620, and 0.460 (chance level = 0.2), averagely 0.540 ± 0.092. For the other seven models (NB, KNN, Ada, RF, SVM, RNN, and GRU), the average accuracies of the five subjects are 0.322, 0.357, 0.380, 0.433, 0.482, 0.430, and 0.527, respectively. Paired-sample t test shows that the accuracy of the LSTM-based and GRU-based decoding model is significantly higher than that of the five traditional models and the RNN-based method ($p < .05$). We found that the LSTM-based and GRU-based model achieved very close decoding accuracy, because GRU and LSTM share very similar structures of the network.

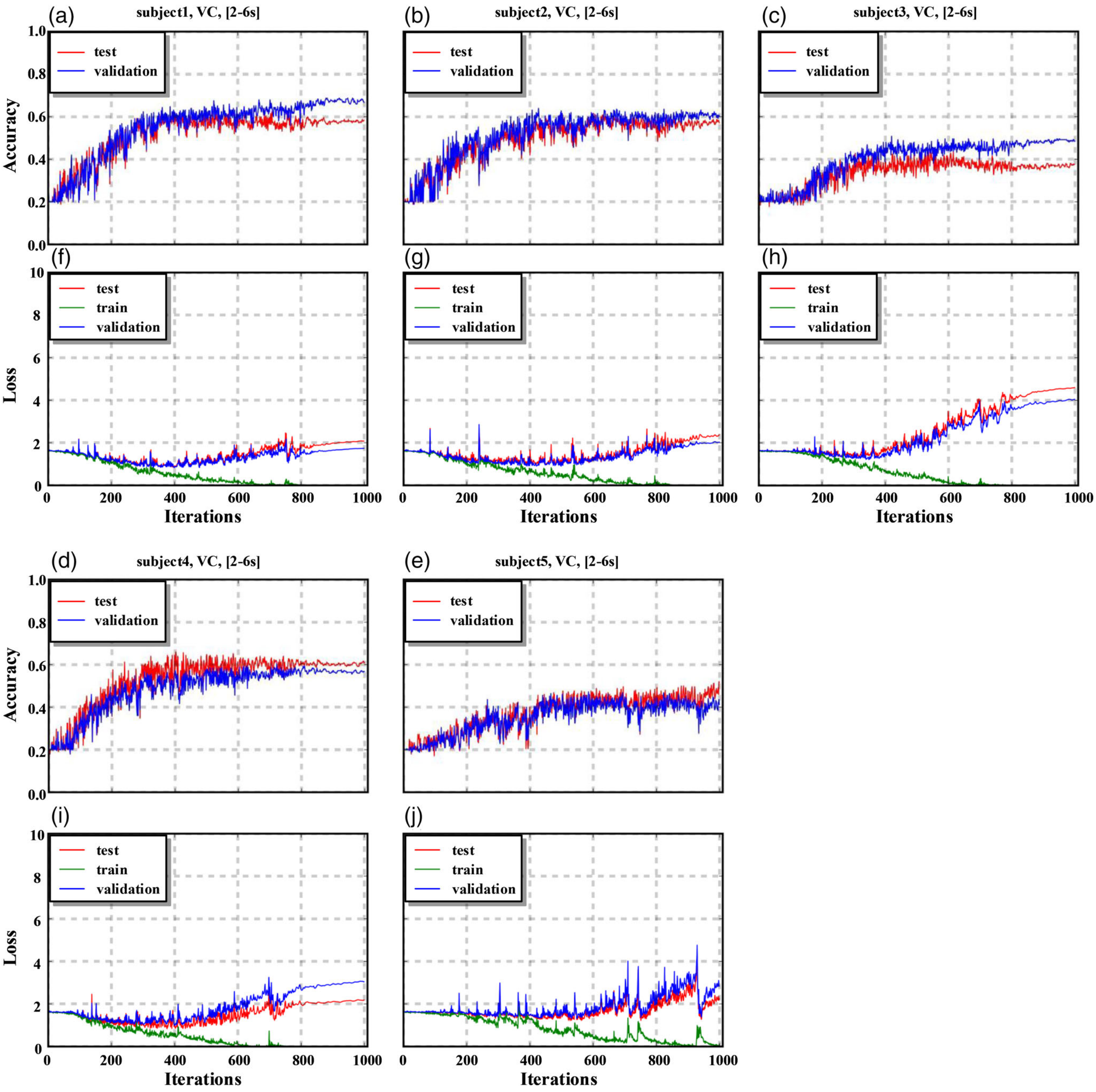


FIGURE 5 The curves of the accuracy and loss over time for the training process with the long short-term memory (LSTM)-based decoding model in the visual cortex (VC) from five subjects. (a–e) The accuracy curve of the five subjects in the iterative training process. The red curve indicates the accuracy of the test data. The blue curve indicates the accuracy of the validation data. (f–j) The corresponding loss of the five subjects in the iterative training process. The red curve, blue curve, and green curve indicate the loss of the test data, validation data, and training data, respectively

4.4 | Comparison of decoding performance of different visual areas

Previous studies have shown a hierarchical correspondence between cortical hierarchy and the levels of visual feature representations (Dong, Wang, & Hu, 2018; Shen, Horikawa, Majima, & Kamitani, 2019). To investigate the function of different visual areas in brain decoding of object categories, we also compared the decoding accuracies of different visual areas.

According to the retinotopic experiment, V1, V2, V3, LVC, HVC, and VC are defined. Here, multitime visual response signals of 2–6 s from different visual areas is put into the LSTM-based decoding model for decoding. The decoding accuracies of the five subjects under different visual areas are shown in Figure 9. The average decoding accuracy of V1, V2, V3, LVC, HVC, and VC from the five subjects is 0.287, 0.300, 0.282, 0.322, 0.514, and 0.540, respectively. Paired-sample *t* test shows that the decoding accuracy of HVC or VC is significantly higher than that of the LVC (V1, V2, V3, or LVC) for all subjects ($p < .05$).

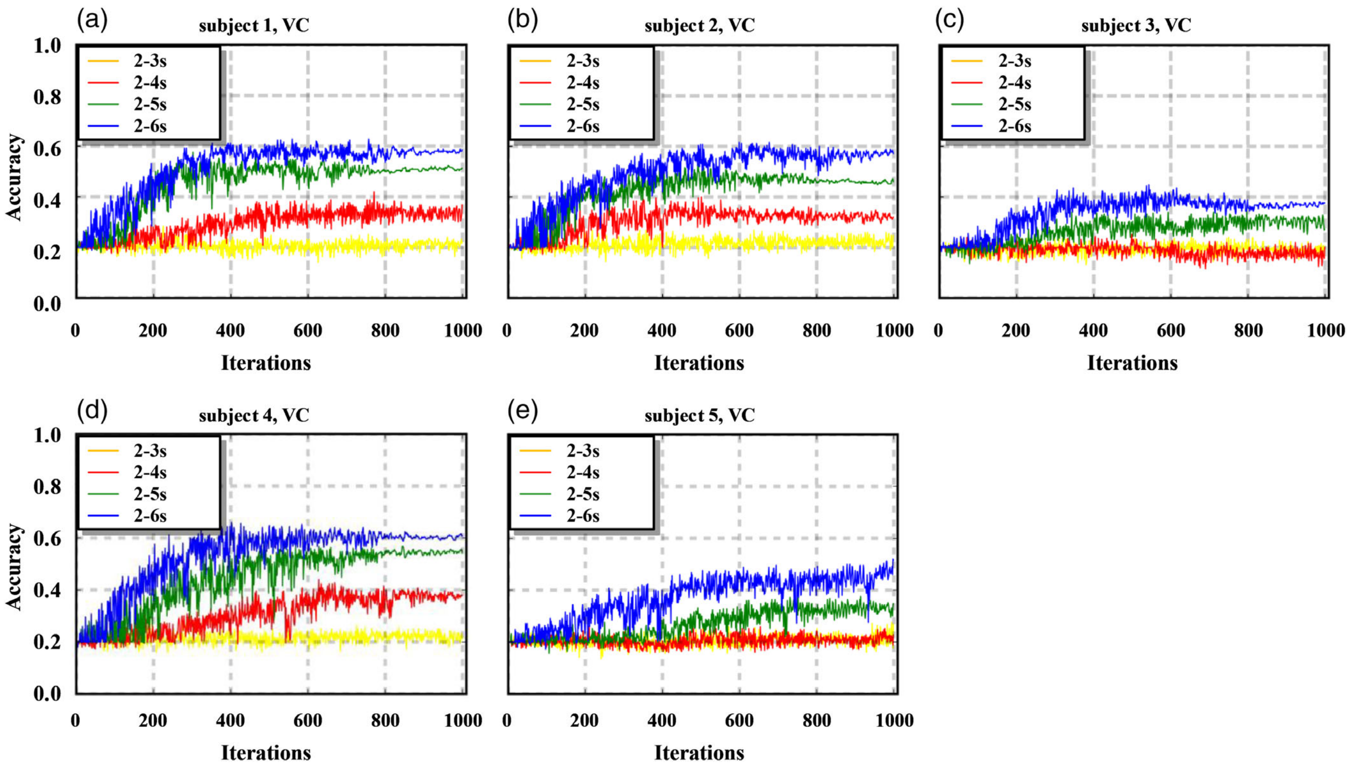


FIGURE 6 Comparison of decoding performance of multitime (second–third, second–fourth, second–fifth, and second–sixth seconds) visual response signals with the long short-term memory (LSTM)-based decoding model in the visual cortex (VC) from five subjects. (a–e) The accuracy curves of the five subjects with different durations in the iterative training process. The yellow curve, red curve, green curve, and blue curve indicate the accuracy of the test data using the second–third, second–fourth, second–fifth, and second–sixth seconds of the visual response signals, respectively

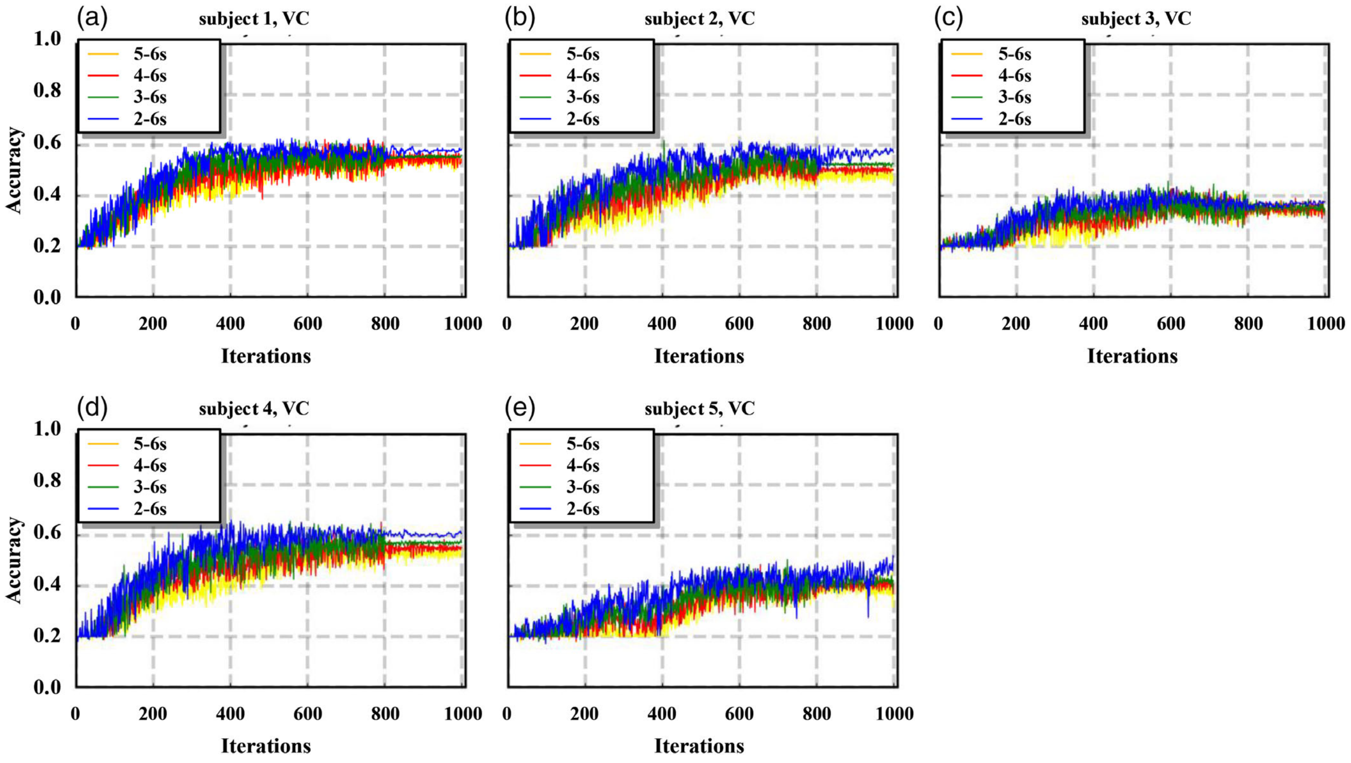


FIGURE 7 Comparison of decoding performance of multitime (fifth–sixth, fourth–sixth, third–sixth, and second–sixth seconds) visual response signals with the long short-term memory (LSTM)-based decoding model in the visual cortex (VC) from five subjects. (a–e) The accuracy curves of the five subjects with different durations in the iterative training process. The yellow curve, red curve, green curve, and blue curve indicate the accuracy of the test data using the fifth–sixth, fourth–sixth, third–sixth, and second–sixth seconds of the visual response signals, respectively

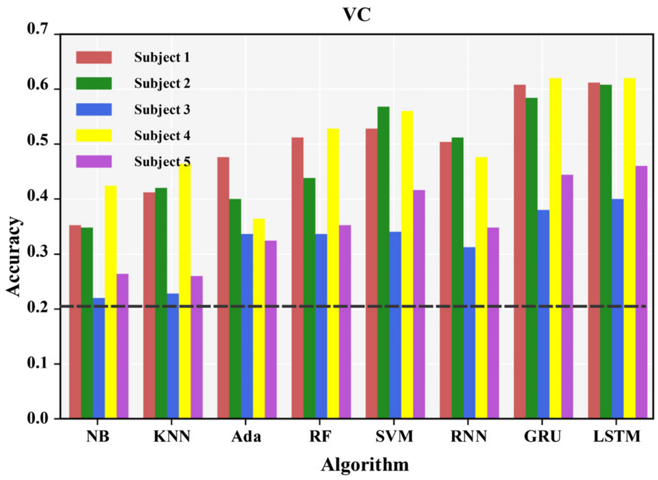


FIGURE 8 Comparison of the decoding accuracy of the eight methods in the visual cortex (VC) from five subjects. The dark dashed line represents the chance level for classifying the five categories

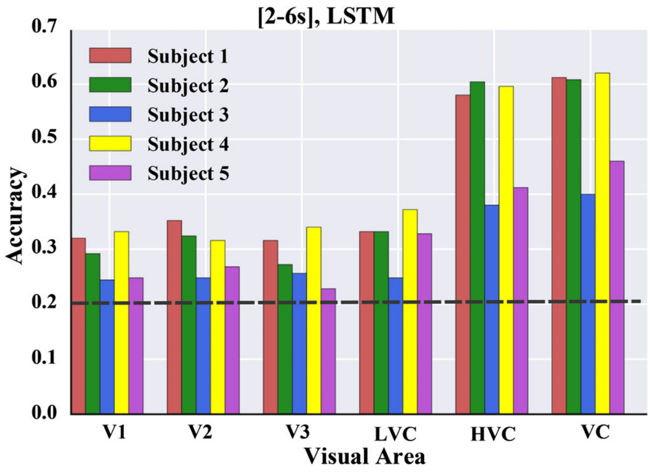


FIGURE 9 Comparison of the decoding accuracy of different visual areas with the long short-term memory (LSTM)-based decoding model from five subjects. The dark dashed line represents the chance level for classifying the five categories

5 | DISCUSSION

Understanding how the human brain perceives the outside world is an important goal of neuroscience (Güçlü & van Gerven, 2017). As noninvasive neural detection technology, both fMRI and EEG have advantages and disadvantages in "brain-reading." fMRI has very good spatial resolution, which means we can obtain neural signals of thousands of voxels for each stimuli. However, it is an indirect effect related to the changes in blood flow that follow the changes in neural activity, showing obvious time dependence (Lin et al., 2009). Although the peak signal of fMRI represents the strongest neural response for the stimuli, the signals before and after the peak still contain lots of stimuli-related neural activities, and they can be useful if properly utilized. Our result in Supplementary Table S2 shows that the single-time signal of 2, 3, 4, and 5 s can also realize category decoding to a certain degree, especially for the fifth second, which is close to the sixth second, showing pretty good decoding accuracy although it is not as good as the sixth second. Therefore, making full use of the multitime fMRI signals rather than the peak ones can help improve decoding performance. The comparisons of multitime response signals in Figures 6 and 7 provide further evidences that 2–6 s response signals which covered the whole time course from the appearance of stimuli to the strongest response provide the most semantic category information for decoding, achieving higher decoding performance than single-time or shorter-time signals. These results illustrate the positive effect of multitime fMRI signals in improving the decoding performance. We suggest researchers strive to incorporate the multitime information as much as possible in future fMRI studies.

Compared with traditional models, neural networks (RNN-based, LSTM-based, GRU-based) with temporal sequence simulation capabilities can be used to simulate the dynamic process to capture the time dependence in fMRI data. In this paper, a comprehensive visual decoding analysis is performed from visual response signals of single time and multitime. The research shows that the LSTM-based and GRU-based decoding model can well capture the time dependence of the fMRI signals evoked by visual stimuli, and their decoding accuracy significantly exceeds the traditional models without sequence simulation capabilities. In addition, since GRU has a similar network structure with LSTM, both of them have introduced the gate mechanism, so there is no significant difference in decoding performance. The decoding accuracy of the LSTM or GRU based decoding model is also significantly higher than the RNN-based decoding model. The possible reason is that the gate mechanism introduced in LSTM or GRU can better fit the dynamic process of the time-dependence signals. LSTM and GRU may have an important impetus to the time-dependence decoding research of the brain in the future.

We not only compared the decoding performance of different methods, but also compared the decoding performance of different visual areas. From the multitime visual response signals of different visual areas (including V1, V2, V3, LVC, HVC, VC), we used the LSTM-based decoding model to decode the visual perceptual category. Our result shows that the decoding accuracy of the HVC or VC is significantly better than that of the LVC (V1, V2, V3, or LVC). The result suggests that the HVC plays an important role in decoding task of object category evoked by natural images. Previous research shows that the functional topography of the HVC reflects an organized category-selective map with particular stimulus category eliciting distinct patterns of cortical activation (Dong et al., 2018; Horikawa & Kamitani, 2017; Shen et al., 2019). Through the indicators of decoding accuracy, the results indirectly support the view that the HVC contains more category-selective semantic information than the LVC.

For conclusion, we proposed an LSTM-based decoding model that makes it possible to classify five categories of the visual stimuli from brain activity. There are three main contributions of this study: (a) the time-dependence fMRI signals are used to improve the decoding performance of five categories of natural images stimulus, achieving about 0.54 (the chance is 0.2) accuracy, (b) by comparing the performance of the decoding models constructed by neural nets and traditional methods, we demonstrate that the LSTM-based and GRU-based decoding model is more suitable for processing fMRI data which contains the dynamic response delay, and (c) we proved that the HVC plays more important role in decoding object category evoked by natural stimulus, which is consistent with previous investigation (Horikawa & Kamitani, 2017). In short, LSTM shows advantages in processing the spatiotemporal fMRI signals and may have promising power for helping us decode complex visual experiences.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (61773094, 61573080, 61533006, U1808204, 31730039, 31671133, and 61876114); the Ministry of Science and Technology of China (2015CB351701); National Major Scientific Instruments and Equipment Development Project (ZDYZ2015-2); and Chinese Academy of Sciences Strategic Priority Research Program B grants (XDB32010300).

CONFLICT OF INTEREST

The authors declared that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

DATA AVAILABILITY STATEMENT

When the paper is received, we will open the original data.

ORCID

Jiang Zhang https://orcid.org/0000-0002-0783-3705
Huafu Chen https://orcid.org/0000-0002-4062-4753

REFERENCES

Auerbach, E. J., Xu, J., Yacoub, E., Moeller, S., & Uğurbil, K. (2013). Multiband accelerated spin-echo echo planar imaging with reduced peak RF power using time-shifted RF pulses. *Magnetic Resonance in Medicine*, *69*, 1261–1267.

Barragan-Jason, G., Cauchoix, M., & Barbeau, E. (2015). The neural speed of familiar face recognition. *Neuropsychologia*, *75*, 390–401.

Behroozi, M., & Daliri, M. R. (2014). Predicting brain states associated with object categories from fMRI data. *Journal of Integrative Neuroscience*, *13*, 1–23.

Behroozi, M., Daliri, M. R., & Shekarchi, B. (2015). EEG phase patterns reflect the representation of semantic categories of objects. *Medical & Biological Engineering*, *54*, 1–17.

Brewer, A. A., Liu, J., Wade, A. R., & Wandell, B. A. (2005). Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nature Neuroscience*, *8*, 1102–1109.

Burkhalter, A., Felleman, D., Newsome, W., & van Essen, D. (1986). Anatomical and physiological asymmetries related to visual areas V3 and VP in macaque extrastriate cortex. *Vision Research*, *26*, 63–80.

Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., & Turret, J. (2011). High temporal resolution decoding of object position and category. *Journal of Vision*, *11*, 9–9.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*, 704–717.

Chen, Y.-W., & Lin, C.-J. (2006). *Combining SVMs with various feature selection strategies. Feature extraction* (pp. 315–324). Berlin, Heidelberg: Springer.

Contini, E. W., Wardle, S. G., & Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia*, *105*, 165–176.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)"brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*, 261–270.

Cox, K. M., & Kable, J. W. (2014). BOLD subjective value signals exhibit robust range adaptation. *Journal of Neuroscience*, *34*, 16533–16543.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Paper presented at the 2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Miami Beach, FL.

Dong, Q., Wang, H., & Hu, Z. (2018). Commentary: Using goal-driven deep learning models to understand sensory cortex. *Frontiers in Computational Neuroscience*, *12*, 4.

Dougherty, R. F., Koch, V. M., Brewer, A. A., Fischer, B., Modersitzki, J., & Wandell, B. A. (2003). Visual field representations and locations of visual areas V1/2/3 in human visual cortex. *Journal of Vision*, *3*, 1–1.

Dumoulin, S. O., & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, *39*, 647–660.

Fishman, R. S. (1997). Gordon Holmes, the cortical retina, and the wounds of war. *Documenta Ophthalmologica*, *93*, 9–28.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, *38*, 11–21.

Güçlü, U., & van Gerven, M. A. (2017). Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in Computational Neuroscience*, *11*, 7.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., & van Gerven, M. A. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. In *Advances in neural information processing systems 30* (pp. 4246–4257). Long Beach, CA: NIPS.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2430.

Haynes, J.-D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*, 686–691.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*, 15037.

Horikawa, T., Tamaki, M., Miyawaki, Y., & Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science*, *340*, 639–642.

Huang, W., Yan, H., Liu, R., Zhu, L., Zhang, H., & Chen, H. (2018). F-score feature selection based Bayesian reconstruction of visual image from human brain activity. *Neurocomputing*, *316*, 202–209.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, *28*, 229–289.

Jafakesh, S., Jahromy, F. Z., & Daliri, M. R. (2016). Decoding of object categories from brain signals using cross frequency coupling methods. *Biomedical Signal Processing and Control*, *27*, 60–67.

Jahromy, F. Z., & Daliri, M. R. (2017). Semantic category-based decoding of human brain activity using a Gabor-based model by estimating intracranial field potential range in temporal cortex. *Journal of Integrative Neuroscience*, *16*, 419–428.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*, 679–685.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*, 352–355.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ariXv*. https://arxiv.org/abs/1412.6980v9.

Lin, A.-L., Fox, P. T., Yang, Y., Lu, H., Tan, L.-H., & Gao, J.-H. (2009). Time-dependent correlation of cerebral blood flow with oxygen metabolism in activated human visual cortex as measured by fMRI. *NeuroImage*, *44*, 16–22.

Lippert, M. T., Steudel, T., Ohl, F., Logothetis, N. K., & Kayser, C. (2010). Coupling of neural activity and fMRI-BOLD in the motion area MT. *Magnetic Resonance Imaging*, *28*, 1087–1094.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-A., Morito, Y., Tanabe, H. C., … Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, *60*, 915–929.

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., & Uğurbil, K. (2010). Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine*, *63*, 1144–1153.

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, *105*, 215–228.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–430.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical parametric mapping: The analysis of functional brain images*, London: Elsevier.

Polat, K., & Güneş, S. (2009). A new feature selection method on classification of medical datasets: Kernel F-score feature selection. *Expert Systems with Applications*, *36*, 10367–10373.

SamPenDu, D. (2017). SamSrf toolbox for pRF mapping.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience Biobehavioral Reviews*, *42*, 9–34.

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS Computational Biology*, *15*, e1006633.

Song, S., Zhan, Z., Long, Z., Zhang, J., & Yao, L. (2011). Comparative study of SVM methods combined with voxel selection for object category classification on fMRI data. *PLoS One*, *6*, e17191.

Szaflarski, J. P., Difrancesco, M., Hirschauer, T., Banks, C., Privitera, M. D., Gotman, J., & Holland, S. K. (2010). Cortical and subcortical contributions to absence seizure onset examined with EEG/fMRI. *Epilepsy & Behavior*, *18*, 404–413.

Tafreshi, T. F., Daliri, M. R., & Ghodousi, M. (2019). Functional and effective connectivity based features of EEG signals for object recognition. *Cognitive Neurodynamics*, *13*, 555–566.

Taghizadeh-Sarabi, M., Daliri, M. R., & Niksirat, K. S. (2015). Decoding objects of basic categories from electroencephalographic signals using wavelet transform and support vector machines. *Brain Topography*, *28*, 33–46.

Torabi, A., Zareayan Jahromy, F., & Daliri, M. R. (2017). Semantic category-based classification using nonlinear features and wavelet coefficients of brain signals. *Cognitive Computation*, *9*, 702–711.

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, *78*, 1550–1560.

Yang, Z., Huang, Z., Gonzalez-Castillo, J., Dai, R., Northoff, G., & Bandettini, P. (2014). Using fMRI to decode true thoughts independent of intention to conceal. *NeuroImage*, *99*, 80–92.

Zong, X., Kim, T., & Kim, S.-G. (2012). Contributions of dynamic venous blood volume versus oxygenation level changes to BOLD fMRI. *NeuroImage*, *60*, 2238–2246.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Huang W, Yan H, Wang C, et al. Long short-term memory-based neural decoding of object categories evoked by natural images. *Hum Brain Mapp*. 2020; 41:4442–4453. https://doi.org/10.1002/hbm.25136