Neural Networks 109 (2019) 19–30

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

The importance of recurrent top-down synaptic connections for the anticipation of dynamic emotions

Martial Mermillod [a,b,*], Yannick Bourrier [c,d], Erwan David [e,f], Louise Kauffmann [g], Alan Chauvin [a], Nathalie Guyader [g], Frédéric Dutheil [h,i], Carole Peyrin [a]

[a] University Grenoble Alpes, University Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France
[b] University Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[c] University Pierre & Marie Curie, LIP6, F-75005 Paris, France
[d] University Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[e] Université de Nantes, F-44000 Nantes, France
[f] LS2N UMR CNRS 6004, F-44000 Nantes, France
[g] University Grenoble Alpes, CNRS, Grenoble INP, GIPSA-LAB, F-38000 Grenoble, France
[h] Université Clermont Auvergne, CNRS, LaPSCo, Physiological and psychosocial stress, University Hospital of Clermont-Ferrand, CHU Clermont-Ferrand, Preventive and Occupational Medicine, WittyFit, F-63000 Clermont-Ferrand, France
[i] Australian Catholic University, Faculty of Health, Melbourne, Victoria, Australia

ARTICLE INFO

Article history:
Received 9 January 2018
Received in revised form 8 September 2018
Accepted 11 September 2018
Available online 9 October 2018

Keywords:
Predictive brain
Neural network modeling
Emotional facial expressions
Top-down processes

ABSTRACT

Different studies have shown the efficiency of a feed-forward neural network in categorizing basic emotional facial expressions. However, recent findings in psychology and cognitive neuroscience suggest that visual recognition is not a pure bottom-up process but likely involves top-down recurrent connectivity. In the present computational study, we compared the performances of a pure bottom-up neural network (a standard multi-layer perceptron, MLP) with a neural network involving recurrent top-down connections (a simple recurrent network, SRN) in the anticipation of emotional expressions. In two complementary simulations, results revealed that the SRN outperformed the MLP for ambiguous intensities in the temporal sequence, when the emotions were not fully depicted but when sufficient contextual information (related to previous time frames) was provided. Taken together, these results suggest that, despite the cost of recurrent connections in terms of energy and processing time for biological organisms, they can provide a substantial advantage for the fast recognition of uncertain visual signals.

1. Introduction

One of the most important characteristics of living organisms is not only having the capacity to anticipate perceptual events but also to detect the emotional content of the environment. To this end, the ability to use knowledge from past experiences to anticipate future events is vital. Several studies in cognitive neuroscience have recently provided experimental evidence supporting this view (Bullier, 2001; Kauffmann, Ramanoël, & Peyrin, 2014; O'Callaghan, Kveraga, Shine, Adams, & Bar, 2017; Summerfield & Egner, 2009; Trapp & Bar, 2015). According to this theoretical approach, visual recognition is not the result of pure bottom-up processes from the perceptual system (e.g. the retina, the lateral geniculate nucleus and the occipital cortex) to the high-level cortical areas dedicated to visual cognition (e.g. identification or categorization of stimuli or events). Instead, top-down processes might be rapidly triggered by the fast processing of a magnocellular signal (e.g. low spatial frequencies) in high-level associative cerebral areas, and especially in the orbitofrontal cortex (OFC; Bar et al., 2006; Kauffmann, Bourgin, Guyader, & Peyrin, 2015; Kauffmann, Chauvin, Pichat, & Peyrin, 2015; Kveraga, Boshyan, & Bar, 2007). This signal may then be transferred to pre-activate lower order perceptual or associative areas that ultimately mediate visual recognition and visual categorization (i.e. the inferotemporal cortex processing a larger band of the spatial frequency spectrum, including high-spatial frequency information). This process is believed to increase the efficiency of low-level recognition and categorization areas through the pre-activation of perceptual expectations about the visual input (Bar, 2004; Beffara et al., 2015). The first neurobiological evidence in favor of this model was provided by magnetoencephalography (MEG) experiments (Bar et al., 2006) that pointed to an activation in the OFC (about 130 ms after

* Corresponding author at: University Grenoble Alpes, Laboratoire de Psychologie et NeuroCognition, BP 47, 38040 Grenoble Cedex 9, France.
E-mail address: Martial.Mermillod@univ-grenoble-alpes.fr (M. Mermillod).

https://doi.org/10.1016/j.neunet.2018.09.007
0893-6080/

stimulus onset) before the inferior temporal cortex (about 180 ms after stimulus onset) during single object recognition. Moreover, recent functional magnetic resonance imaging (fMRI) studies investigating the effective connectivity between these regions using Dynamic Causal Modeling (DCM) have revealed that the magnocellular signal increased the connectivity strength from the OFC to the inferotemporal cortex (Kauffmann, Bourgin et al., 2015; Kauffmann, Chauvin et al., 2015; Kveraga, Boshyan et al., 2007).

With regard to emotional processes, it was assumed that this top-down neural mechanism could be efficient during emotion regulation (Barrett & Bar, 2009; Beffara et al., 2015; Kawasaki et al., 2001; Kveraga, Ghuman, & Bar, 2007; Mermillod, Droit-Volet, Devaux, Schaefer, & Vermeulen, 2010). Given the involvement of the OFC in anticipating visual information and evaluating emotional information (Kawasaki et al., 2001; Niedenthal, Mermillod, Maringer, & Hess, 2010), we could assume a link, at a functional level, between the prediction of events in temporal sequences and the anticipation of dynamic emotional stimuli. The use of dynamic facial expressions thus constitutes a type of stimuli particularly appropriate for addressing this question at the perceptual categorization level and in terms of disentanglement between ambiguous EFEs (in a continuum from neutral to the APEX of different EFEs).

In terms of neural computation, among the different techniques used to process temporal sequence prediction, Simple Recurrent Networks (Elman, 1990) constitute a type of neural network that uses neural connectivity from hidden (or "associative") layers to input (or "perceptual") layers in order to predict temporal sequences. Albeit very simplified compared to a biological neural system, this type of artificial neural network can learn very complex structures over time (Bengio, Ducharme, Vincent, & Jauvin, 2003).

However, we have to notice the striking differences between biological systems such as the human brain that greatly use recurrent connectivity (Bullier, 2001; Sherman & Guillery, 2002), and state-of-the-art neural network modeling that shows outstanding performance based on pure bottom-up processes (i.e. from perceptual to associative areas but without any recurrence from associative to perceptual areas). These models are efficient and occasionally more effective at recognizing faces than humans, even in noisy environments (Li, Lin, Shen, Brandt, & Hua, 2015; Parkhi, Vedaldi, & Zisserman, 2015; Taigman, Yang, Ranzato, & Wolf, 2014; Wang & Cottrell, 2016, 2017; Wang, Gauthier, & Cottrell, 2016). Similarly, more simple and bottom-up neural networks, such as MultiLayer Perceptrons (MLP), are sufficient to match or even exceed human performance levels when recognizing static images of emotional facial expressions (Dailey & Cottrell, 1999; Dailey, Cottrell, Padgett, & Adolphs, 2002; Mermillod, Bonin, Mondillon, Alleysson, & Vermeulen, 2010; Mermillod, Droit-Volet et al., 2010; Mermillod, Vermeulen, Lundqvist, & Niedenthal, 2009; Mermillod, Vuilleumier, Peyrin, Alleysson, & Marendaz, 2009; Tong, Joyce, & Cottrell, 2008). If high accuracy is possible using only bottom-up processes from perceptual to cognitive areas, the question is to determine what is the advantage for biological systems in using recurrent top-down processes when recognizing emotional events such as dynamic emotional expressions.

It is important to specify that our current study does not provide a precise simulation of Bar's model (2004). This model involves different brain regions with different spatial and temporal frequency specificities, in addition to phase synchronization between frontal and temporal regions (Bar et al., 2006). A precise brain-inspired neural network modeling of this model goes far beyond the scope of the current paper. The goal of the present article is more basic, but also more general, namely: could recurrent connectivity from associative to perceptual areas be useful for the categorization of dynamic emotional events? Our hypothesis is that recurrent neural connectivity provides an advantage when recognizing and anticipating more ambiguous emotional signals. For instance, at the beginning of a sequence consisting of neutral to higher intensity emotional expressions. To validate this very general hypothesis using computational models, we compared the most simple and most comparable types of neural networks in order to test the importance of recurrent connections, with everything else being as similar as possible (i.e. identical learning rate, synaptic weight correction, training/test procedure, etc.) Namely, we compared a simple Recurrent Neural Network (SRN) to a simple Multi-Layer Perceptron (MLP) for the correct prediction of dynamic emotions.

Compared to artificial neural networks, biological neural systems often deal with complex, subtle, and dynamic visual stimuli which need to be anticipated rapidly for survival purposes. Therefore, in two complementary simulations, both networks received images extracted from videos as inputs. These images showed faces gradually varying from neutral to six different emotional expressions (anger, disgust, fear, happiness, sadness, and surprise). As we expect an advantage for the recurrent neural connectivity, two results should be observed. Firstly, the SRN should be more efficient than the MLP during the earliest stages of the video when the signal (i.e. the emotional expression) is not well differentiated and ambiguous (i.e. close to neutral). Secondly, the SRN should outperform the MLP, especially for subtle EFEs involving less distinctive features, such as fear or disgust, compared to other, less ambiguous EFEs (e.g. happiness, according to Ekman and Friesen (1976).

2. Simulation 1

2.1. Method

The following simulation was performed to estimate the impact of recurrent connectivity from associative to perceptual layers when predicting which emotion is going to be displayed. According to the proactive brain hypothesis (Bar, 2007), top-down recurrent connections provide cues to guide perception of what should be expected next. To provide further evidence for this hypothesis, we compared the performances of two different neural networks (SRN vs MLP). Note that the simplest and more parsimonious core neural networks were used to ensure that the differences were not due to hyper-parameters, algorithms or training procedures but only to recurrent connectivity. The first classifier was a MLP using the standard back-propagation training algorithm, as used in Mermillod, Bonin et al. (2010) and Mermillod, Droit-Volet et al. (2010). The second classifier, an SRN, was the same MLP to which a reinjection of the hidden layer was added at the level of the perceptual units. The procedure was very similar to the original SRN proposed by Elman (1990): at time t, the input units (from the perceptual part of the input layer) received the first input of the sequence (i.e., the first frames of the video) whereas the context units were set to initial random values. Both input units and context units were fed-forward to the hidden units. Then the hidden units were fed forward to the output units while the hidden units were also copied as identical to the context units. During the modification of the synaptic weights, the observed output was compared with the expected output, and the back-propagation algorithm was applied to adjust the synaptic weights across the entire neural network (including input and context units). Therefore, the SRN differed from the MLP in that its hidden layer was re-injected as an input to the next pattern, which enabled the network to keep in memory the structure of the visual environment through time.

Preprocessing and stimuli

We used similar stimuli (emotional stimuli) and preprocessing (using Gabor filters) as the EMPATH model (Dailey et al., 2002). The training was done on six facial expressions of primary emotions (anger, disgust, fear, happiness, sadness, and surprise). To test our models in a similar way to EMPATH but on realistic videos, we

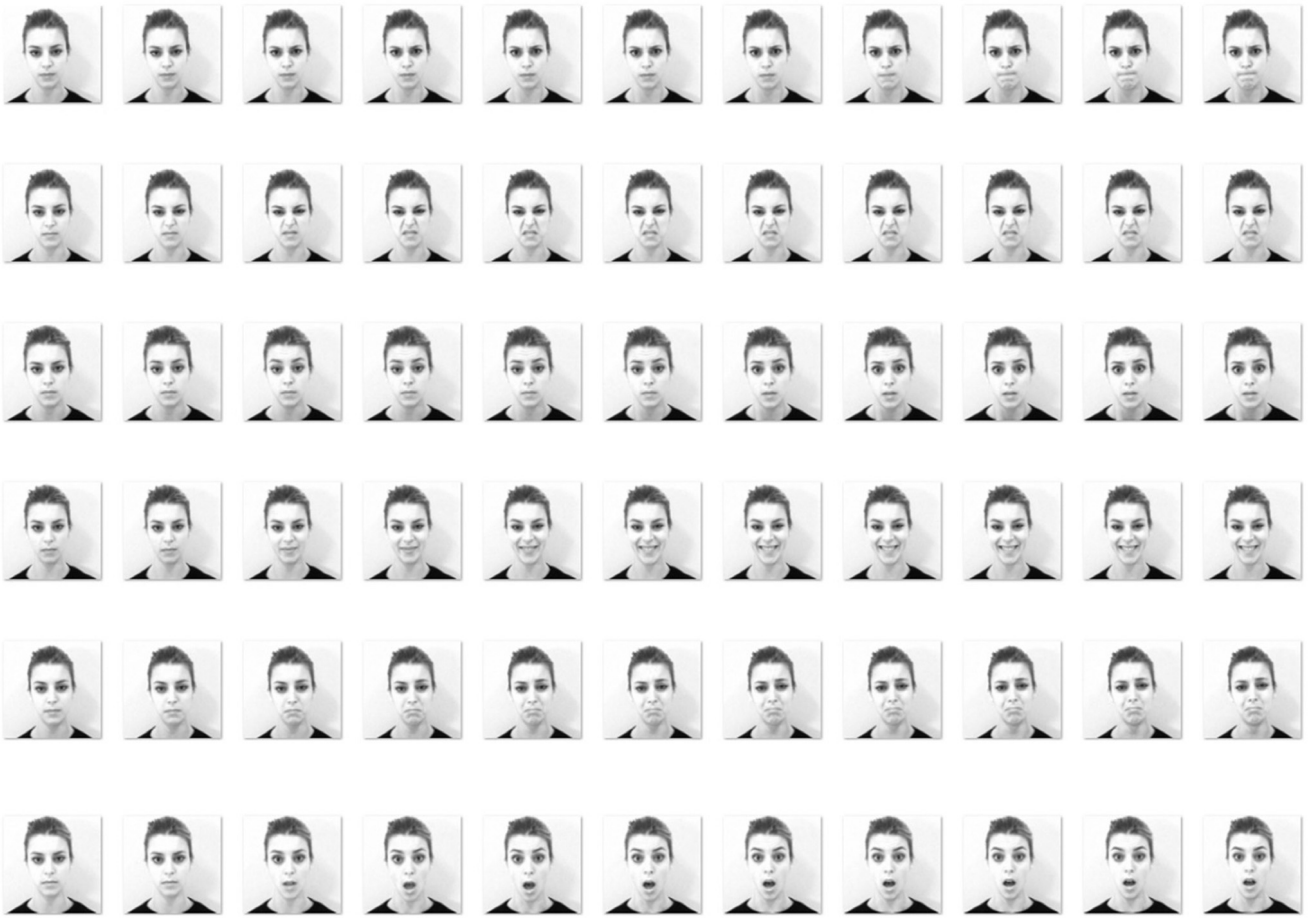

Fig. 1. Example of emotional facial expressions from neutral to the apex of each emotion used for Simulation 1. Training was done sequentially until the tenth intensity (the last intensity does not predict anything).

trained our neural networks with the Amsterdam Dynamic Facial Expression Set (ADFES database, Van Der Schalk, Hawk, Fischer, & Doosje, 2011). The ADFES database provides a more natural display of facial expressions as the frames were extracted from real time videos, while the POFA database (used by Dailey et al., 2002) has actors instructed to perform and maintain strong facial descriptors of primary emotions, based on the Facial Actions Coding System (FACS, Ekman & Friesen, 1976). The database consisted of videos including 7 male and 5 female faces of North European ethnicity. Each video of the database was decomposed into 10 pictures from neutral to the apex of each emotional expression (Fig. 1).

To obtain average information from the videos, we extracted 10 images per video, which corresponded to a static continuum of emotional display ranging from its onset (the first second) to the apex of each EFE. Each image was converted to a 256-level grayscale and cropped from 720 × 576 pixels to a square image of 576 × 576 pixels allowing the removal of the white background around the face. As a result, we obtained a database of 10 images of 12 actors depicting six EFE (i.e. 720 images).

The images were then transferred to the Fourier domain and filtered by a bank of 48 Gabor filters corresponding to six different spatial frequencies, with one octave between the centers of two consecutive spatial frequency channels ($f_i = \{5.41; 10.77; 21.60; 43.20; 86.40; 172.8\}$ cycles per images) and eight different orientations ($\theta = \{0, \pi/8, 2\pi/8, 3\pi/8, 4\pi/8, 5\pi/8, 6\pi/8, 7\pi/8\}$ radians). Different studies have shown that the use of Gabor filters results in a good approximation of the receptive fields of the simple cells of the primary visual cortex (Hubel & Wiesel, 1968), given that the statistical evaluation of the residual error between the difference in the response profiles of V1 simple cells and Gabor filters are not distinguishable from chance (Jones & Palmer, 1987; Jones, Stepnoski, & Palmer, 1987).

The images were transferred in the Fourier domain to improve the speed and simplicity of the mathematical processes and Gabor filters were applied to each thumbnail by means of a multiplication in the spectral domain (which is equivalent to a convolution of the Gabor receptive fields in the spatial domain). The equation of the kernel of Gabor filter in the spectral domain is:

$$G(u, v) = exp\left[-\left(\frac{(u_\theta - f_i)^2}{2\sigma_u^2} + \frac{v_\theta^2}{2\sigma_v^2}\right)\right], \quad (1)$$

with $u_\theta = \mathrm{u}\cos\theta + \mathrm{v}\sin\theta$ and $v_\theta = \mathrm{v}\cos\theta - \mathrm{u}\sin\theta$. σ_u and σ_v are standard deviations of the Gaussian envelop in the u_θ and v_θ direction (i.e. orthogonal to θ).

The outputs of the Gabor filters were the local energy spectra multiplied by the kernel of the Gabor filter. The Gabor filters were applied on subscale images obtained by dividing the original image into a 7 by 7 grid, with each thumbnail vertically and horizontally overlapping the other by 50% (Fig. 2). This resulted in a 48 × 49 = 2352 value vector for each of the 720 images. Overlapping should compensate for the loss of phase information and therefore improve the general performance level of the model (Rumelhart et al., 1985).

We applied this method to each image of the database to feed the perceptual layer (i.e. the input layer) of the neural network. However, given the significant size of the resulting vector, we applied a PCA to reduce data dimensionality by keeping the first 56

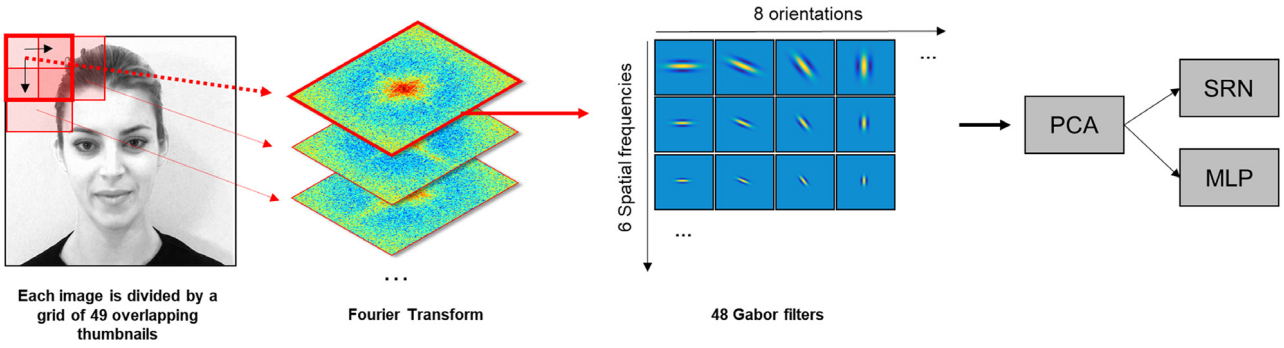

Fig. 2. Each original image (far left) was cut into 49 overlapping thumbnails, which were transferred into the Fourier domain, and to which a bank of 48 Gabor filters were applied. PCA was processed on the resulting values in order to compute a 56-length vector for subsequent SRN or MLP neural networks.

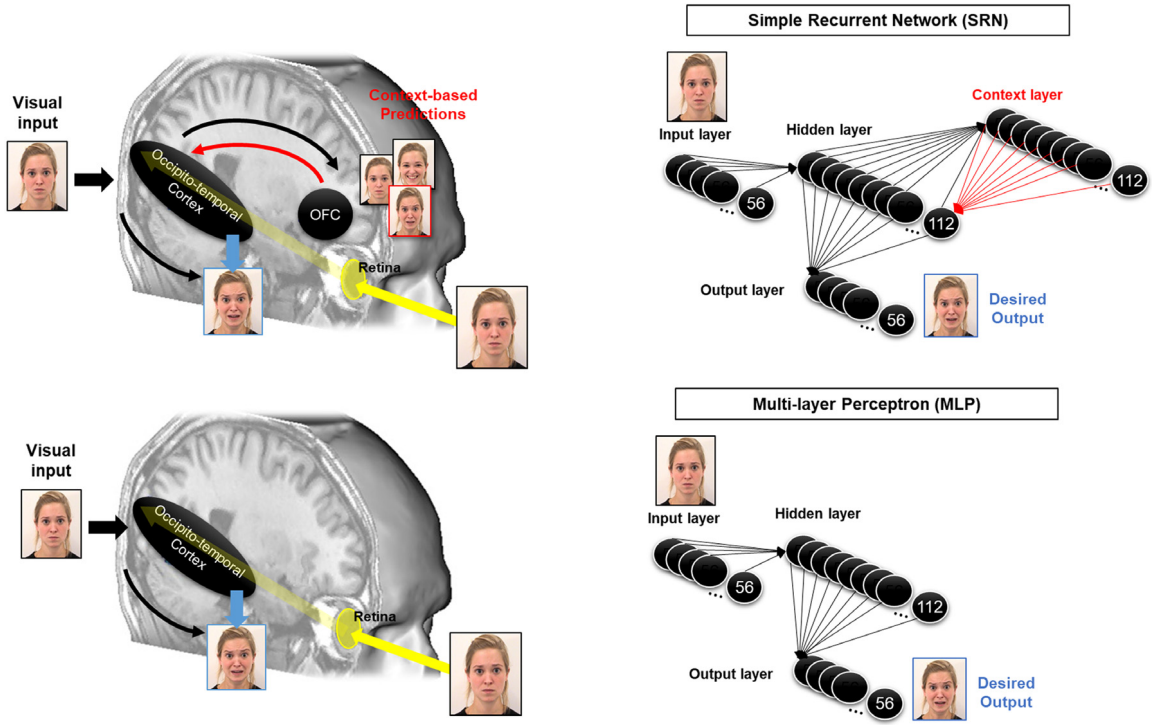

Fig. 3. Visual representation of the recurrent neural network (i.e. the SRN, which simulates top-down synaptic connections from associative to perceptual layers) and the Multi-Layer Perceptron (i.e. the MLP, simulating the bottom-up spreading of activation from perceptual to associative and output layers). Consecutive layers were fully interconnected. The SRN has its hidden layer re-injected as input.

principal components (Dailey et al., 2002). From the resultant matrix of 720 × 2352 value vectors we obtained an array of 720 × 56 eigenvectors explaining about 97% of the variance. Finally, the resultant values were normalized between 0 and 1 across all faces and all emotions, but for each eigenvector independently in order to avoid an over-representation of specific eigenvectors. The PCA then fed into either a MLP or an SRN with 56 input units (constituted by the vector provided by the PCA at time t), 112 hidden layer units and 56 output units (constituted by the vector provided by the PCA at time $t + 1$). We have chosen this number of hidden nodes to stay in range of what was previously used on SRN training when the size of the input and output were the same (Elman, 1990). Therefore, the goal of the neural network was not an auto-association process but was to associate the frame at time (t) with the next frame of the sequence, at time ($t+1$), in order to anticipate its visual environment (i.e. which emotion was the most likely to appear in the next frame of the video sequence). In other words, as we were interested in finding out if the presence of a context layer simulating recurrent brain connections improved EFE categorization, and therefore one's ability to predict which emotion was next to be displayed, the desired outputs for both networks were the next emotional frames, i.e. the next image extracted from the video sequence from which the input image originated. The desired output, also made of 56 units, was then used as the next input in the training procedure, hence simulating the learning of a temporal sequence (Fig. 3). Therefore, both networks were trained in exactly the same way except that we added recurrent connections from hidden to perceptual layers in the SRN. We chose to use very simple MLP algorithms rather than more complex deep neural networks in order to allow a fair and parsimonious comparison between a pure bottom-up (i.e. MLP) process and a top-down neural network (i.e. SRN).

The learning rate was set to 0.1, the momentum to 0.9 and Fahlman offset to 0.1. We used a standard sigmoidal transfer function given by:

$$f(a) = \frac{1}{1 + e^{-a}}, \tag{2}$$

where: *f(a)* is the output activation value and *a* is the sum of the input activation vector multiplied by the input-to-hidden weight matrix.

Given the continuous values of input/output neurons, the error signal was given by the standard Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}, \quad (3)$$

where n is the number of elements in the vector (56), y the expected output, and ŷ the output produced as a result of a forward pass.

Training-test procedure

Both networks were trained in the same way. For each run of the simulation, the faces of 11 out of the 12 actors were randomly selected for training and the last one was kept out of the training phase and was used during the validation phase to test the generalization performance of the neural network. This procedure allowed us to extract the maximum of statistical regularities among the training set given the restricted number of the stimuli in the database. However, in order to ensure a stable average on a reliable test set from the database, this randomized training/test procedure was replicated across 50 runs to ensure that each stimulus was presented a sufficient number of times. This training/test procedure was independently applied for each run. The six EFEs were learnt by the neural network, so there was a total 660 items ($6 \times 11 \times 10$) used for learning (with, respectively: emotional expressions $\times$ actors $\times$ intensities), with intensity being defined here as the extent to which a specific emotion was expressed on the face of the expresser. Low intensities account for neutral emotion displays, whereas high emotion intensities exhibit facial features near or beyond emotional apex. For each iteration, the neural network was trained on a whole sequential order to learn to anticipate the next frame. The EFEs were learnt in a random order across the different iterations constituting a run. The learning phase consisted of over 1000 iterations for the two networks. Synaptic weights were randomized at the beginning of each run to initialize the neural networks before each new training/test procedure. Back-propagation for the MLP, and back-propagation and copy of the context nodes for the SRN were turned off when the last emotional intensity was reached to ensure that only the desired emotional sequences were learned for each iteration. Therefore the only temporal sequences learned were the displays of a participant's facial expression from onset to peak. Output error was obtained by computing the Euclidean distance between the observed output produced by the neural network and the six different $(t+1)$ image vectors, one for each emotion displayed by the same participant. We then used a "winner-take-all" procedure, choosing the smallest Euclidean distance out of the six to determine into which category of emotion the neural network had classified the image.

2.2. Results

Anticipation accuracy of the next frame

In order to investigate the anticipation accuracy of both types of neural network, the observed output values were sorted as true (if the smallest of the 6 Euclidean distances corresponded to the desired emotion) or false (if any other emotion dominated). For each type of neural network, 50 simulations were run with one binary outcome value retrieved for each emotion (anger, disgust, fear, happiness, sadness, and surprise) and intensity (between 1 and 9). A generalized linear mixed model (GLMM, binomial distribution) was created to account for the mixed-effects of this study and the non-Gaussian distribution (Bernouilli distribution) of the data. The model represents the interaction between the two neural network types as a between-factor variable and emotions and intensities as within-factor variables. Primary and interaction effects were explored by means of the Wald Chi-square test, an analysis of the deviance test more suited to studying binary logistic regression mixed models. Post-hoc analyses, when performed, were calculated according to the least square means method with Holm–Bonferroni correction.

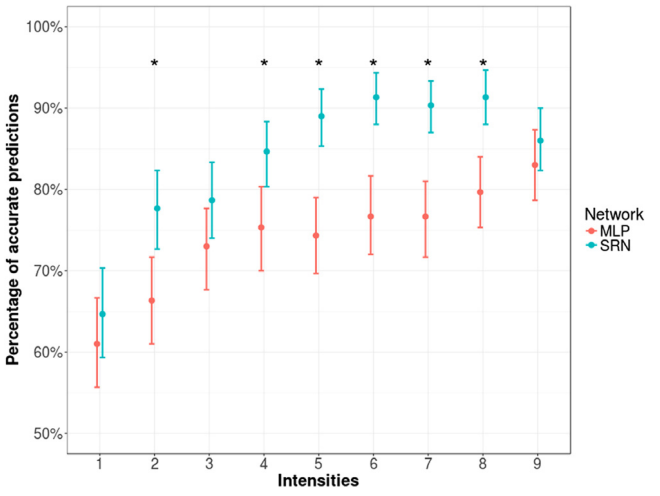

Fig. 4. Mean proportion of correctly anticipated emotions with each type of neural network as a function of the emotional image intensity. Error bars represent standard errors.

The results revealed a main effect of the neural network type ($\chi^2(1, N = 100) = 7.77$, $p < 0.01$) with greater accuracy for the SRN compared to the MLP ($M_{MLP} = 0.74$, $M_{SRN} = 0.84$). A main effect of the intensity ($\chi^2(8, N = 100) = 167.57$, $p < 0.001$) was also observed and suggested that the prediction performances improved as the intensity shifted toward the apex. Moreover, an interaction between the type of neural network and the intensity was also found ($\chi^2(8, N = 100) = 27.27$, $p < 0.001$), suggesting that the advantage of the SRN over the MLP was dependent on the intensity of the facial expression. Pairwise comparisons between SRN and MLP at each of the intensities indicated that the effect was most pronounced for intermediate to high intensities (Fig. 4).

Moreover, a main effect of the emotion ($\chi^2(5, N = 100) = 128.72$, $p < 0.001$) revealed that all emotions were not equally recognized (Fig. 5). The interaction between the type of neural network and the emotion was not significant, but pairwise comparisons revealed that for all emotions, with the exception of happiness, the predictions of the SRN model outperformed those of the MLP model (Fig. 5).

Quadratic error on predictive performance

In addition to the anticipation accuracy of the neural networks, we investigated the ability of both types of neural network to predict the expected values of output neurons at time $t + 1$ after exposure to emotional expressions at time *t*. A Linear Mixed Model (LMM) analysis was created in order to account for the mixed-effect of this study. In the same manner, this new model represents the interaction between the neural network types as a between-factor condition and the within-factor characteristics of emotions and intensities. The Wald Chi-square test was used to retrieve primary and interaction effects, and Holm–Bonferroni corrections were also applied for post-hoc comparisons.

Concerning the continuous variable "quadratic error", results revealed a marginally significant effect of the type of neural network ($\chi^2(1, N = 100) = 3.44$, $p = 0.06$), with less error in the SRN compared to MLP ($M_{MLP} = 0.74$, $M_{SRN} = 0.84$). Besides

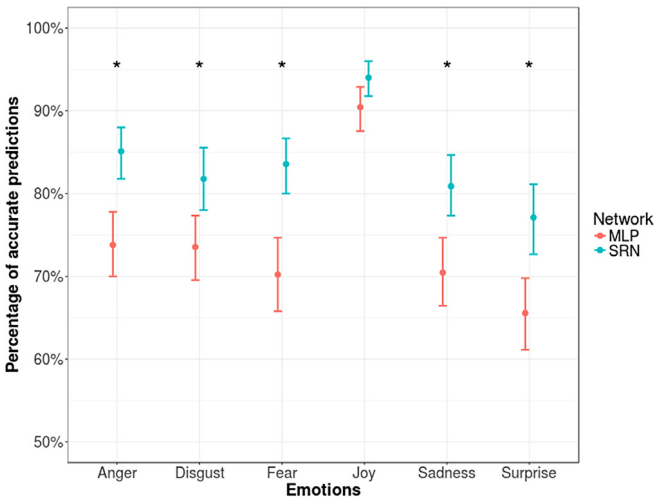

Fig. 5. Mean proportion of correctly anticipated emotions by each type of neural network as a function of the emotional expression. Error bars represent standard errors.

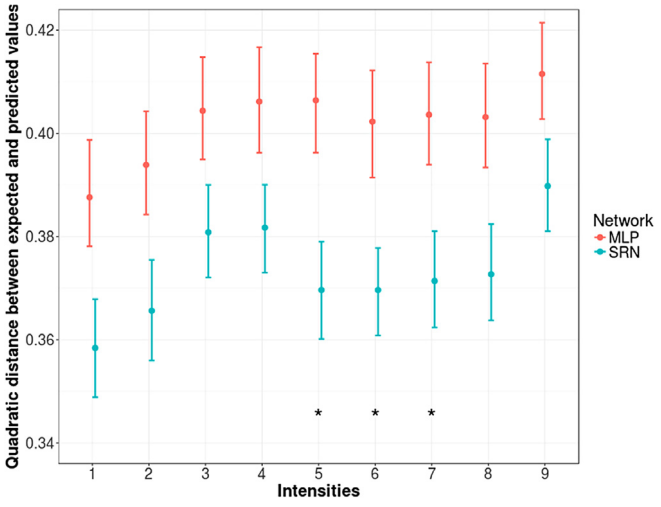

Fig. 6. Mean quadratic error achieved by each type of neural network as a function of the emotional image intensity. Error bars represent standard errors.

this important effect with regard to our initial hypotheses, we observed a significant effect of intensity ($\chi^2(8, N = 100) = 265.02$, $p < 0.001$), with greater error for the intermediate intensities of the emotions, but also a significant effect of emotion ($\chi^2(5, N = 100) = 111.37$, $p < 0.001$), indicating that certain emotions were better predicted than others. More critically, we observed a significant interaction between the type of neural network and the intensity of the emotion ($\chi^2(8, N = 100) = 24.46$, $p < 0.01$). Pairwise comparisons between SRN and MLP at each intensity indicated that the lower prediction error produced by the SRN compared to the MLP is related to the intermediate intensities of the facial expressions (Fig. 6).

We also observed a significant interaction effect between the type of neural network and emotion ($\chi^2(5, N = 100) = 73.73$, $p < 0.001$). Pairwise comparisons revealed that the SRN produced less prediction errors for anger, fear and happiness (Fig. 7).

3. Simulation 2

The aim of Simulation 2 was (i) to test the generalizability of the main results of Simulation 1 on a new database of emotional

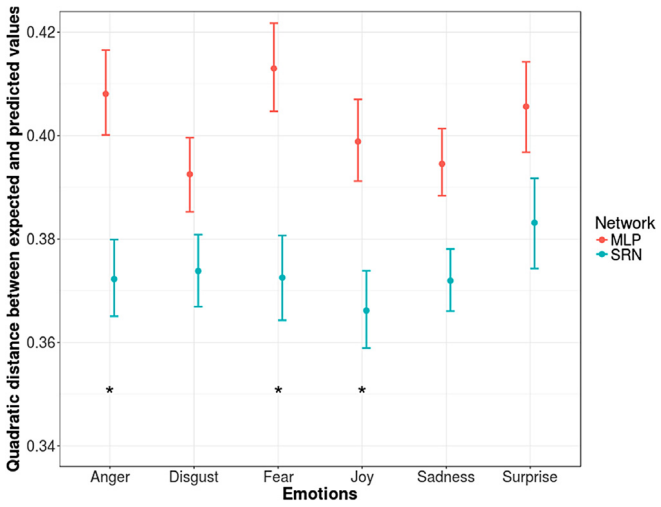

Fig. 7. Mean quadratic error achieved by each type of neural network as a function of the emotional expression. Error bars represent standard errors.

facial expressions, (ii) to remove the arbitrary choice of a given set of eigenvectors following the PCA used in Simulation 1 and (iii) compare the performance of the SRN to a MLP with the same input size. We also aimed to compare the performance of the SRN to a MLP on the basis of an extended version of the MLP which included the same number of synaptic weights compared to the SRN. This would confirm if the better performance of the SRN was not only due to the larger number of synaptic weights inherent to the recurrent connections of this neural network but also to the "snowballing effect" generated by the top-down information provided by the contextual layer.

3.1. Method

As in Simulation 1, the following simulation was performed to estimate the impact of recurrent connectivity between the associative and perceptual layers with respect to the proactive brain hypothesis. This suggests that top-down recurrent connections are important for the anticipation of visual (and specifically emotional) events (Bar, 2007). In order to provide evidence for the importance of top-down recurrent connections on the capacity of an artificial neural network to anticipate emotional events, we tested the neural network on a new database of dynamic videos. This database is largely used in the scientific literature and is provided by the Pictures Of Facial Affect Set (POFA database; Ekman and Friesen, 1976). The PCA was removed as an intermediate step to compress visual information (and therefore, the arbitrary choice of a specific number of eigenvectors) and was replaced by Gabor filters directly applied to the Fourier transformation of the image. The result was closer to a global perception of the image in terms of spatial frequencies and orientations (Mermillod, Bonin et al., 2010; Mermillod, Droit-Volet et al., 2010). We compared the SRN and the MLP performance to an extended MLP (MLPe) with the same number of neurons and synaptic weights as the SRN.

Preprocessing and stimuli

Training was performed on six facial expressions of primary emotions (anger, disgust, fear, happiness, sadness, and anger). We trained our neural networks on the POFA database that provides pictures of actors instructed to perform and maintain strong facial descriptors of primary emotions, based on the Facial Actions Coding System (FACS, Ekman & Friesen, 1978). The database consisted of six male and four female faces of Caucasian ethnicity. Each

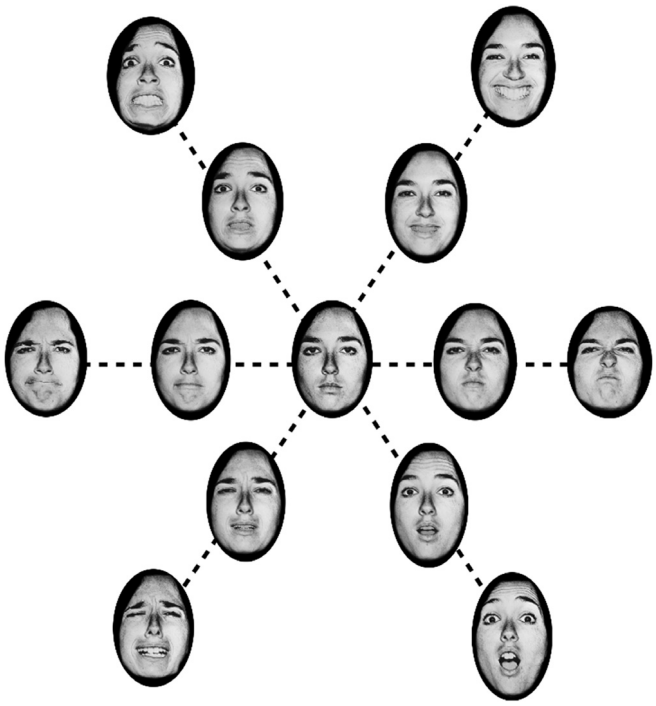

Fig. 8. Example of emotional facial expressions from neutral to the apex of each emotion used for Simulation 2.

emotional sequence was composed of 18 pictures (Fig. 1) by applying a method of morphing between the neutral facial expression and the apex of all emotions. Thus, each emotion intensity range was centered on the neutral expression and shared the same first intensity (0% morphing). The morphing method provides facial expressions that pass the 100% (apex of emotions), reaching 150% and exaggerating the features (see Fig. 8).

As in Simulation 1, facial stimuli were transferred to the Fourier domain and filtered by a bank of 56 Gabor filters corresponding to six different spatial frequencies and 8 different orientations (0, $\pi/8$, $2\pi/8$, $3\pi/8$, $4\pi/8$, $5\pi/8$, $6\pi/8$, $7\pi/8$). The distance between two frequency channel centers was one octave. The difference with Simulation 1 was that we did not use the bank of Gabor filters on a sliding window but directly on the Fourier transform of the entire image. This method avoids the problem of an arbitrary selection of eigenvectors by means of a multiplication of the average energy value in the Fourier domain by the kernel of the Gabor filter at a given location in the spectral (rather than spatial) domain. In other words, each image was encoded in this new simulation by a single bank of 56 Gabor filters. We have previously shown that this method of reducing the input dimensionality remains efficient for the categorization of emotional expressions (Mermillod, Bonin et al., 2010; Mermillod, Droit-Volet et al., 2010).

The architecture of the SRN and MLP were identical to Simulation 1 but we also included an extended version of the MLP which had a similar number of synaptic weight parameters as the SRN:

- MLP: 56 input units (consisting of the vector provided at time t), 112 hidden layer units and 56 output units (consisting of the vector provided at time $t + 1$).
- SRN: 56 input units (consisting of the vector provided at time t) + 112 contextual units (consisting of the neural values of the hidden layer provided at time t − 1), 112 hidden layer units and 56 output units (consisting of the vector provided at time t + 1).
- MLP extended (MLPe): 168 input units (consisting of the concatenation of vectors provided at time t, t + 1 and t + 2), 112 hidden layer units and 56 output units (consisting of the vector provided at time t + 3). During the first three intensities, the input vectors were padded with zeros.

Note that the MLPe has the same number of synaptic parameters as the SRN but is qualitatively processing more information at a time, given that the input sequence was larger (3 frames at a time versus one for the SRN). However, we tested this new situation to ensure that the "snowball effect" generated in the contextual layering of a recurrent still provides more efficient results compared to a large MLP qualitatively processing more information at a time. Artificial neural networks were implemented with *pytorch* (Paszke et al., 2017) and we used the same sigmoidal transfer function as Simulation 1.

Forward passes were performed with batches of all emotions and faces for one intensity, then iterated over intensities by means of forward passes; an Adam optimizer (Kingma & Ba, 2014) was used for gradient descent and updating weight and biases. As in Simulation 1, the loss function was the Root Mean Square Error between the model's output and the expected vector. The SRN's contextual unit was declared as a zero-filled 112-element vector before the forward passes of the first intensity and was updated with the context of the hidden layer for the following intensities. As in Simulation 1, given the continuous values of input/output neurons, we used a RMSE to update the synaptic weights.

Training-test procedure

Both networks were trained with the same procedure as Simulation 1: for each run, 9 out of the 10 faces were selected for training and the last one was kept out of the training phase and was used during the validation phase to test the generalization performance of the neural network. This training/test procedure was applied independently for each run. The six EFEs were learnt by the neural network so that there was a total of $6 \times 9 \times 17 = 918$ items (with emotional expressions × actors × intensities, respectively, as variables) used for learning. For each iteration, the neural network was trained on a whole sequence in a sequential order (in order to learn to "guess" the next frame). The learning phase occurred over 10,000 iterations, which was repeated across 80 runs for the three network types. Synaptic weights were randomized at the beginning of each run to initialize the neural networks before each new training/test procedure. Back-propagation for the different neural networks was turned off when reaching the last emotional intensity to ensure that only the desired emotional sequences were learned for each iteration. Therefore, the only temporal sequences learned were the displays of a participant's facial expression from neutral to peak. The output error was obtained by computing the RMSE between a network's output and the expected output for a given face, emotion and intensity. We then used a "winner-take-all" procedure, choosing the lowest Euclidean distance out of the 6 to determine into which emotional category the neural network had classified the image.

3.2. Results

Anticipation accuracy of the next frame

In order to investigate the anticipation accuracy for each neural network type (MLP, MLPe and SRN), the observed output values were sorted as true if the smallest out of the 6 Euclidean distance corresponded to the target emotion, or false if any other emotion prediction dominated. For each type of neural network, 80 simulations were run; one binary outcome value was retrieved for each combination of emotion (anger, disgust, fear, happiness, sadness, and surprise) and intensity (between 1 and 17).

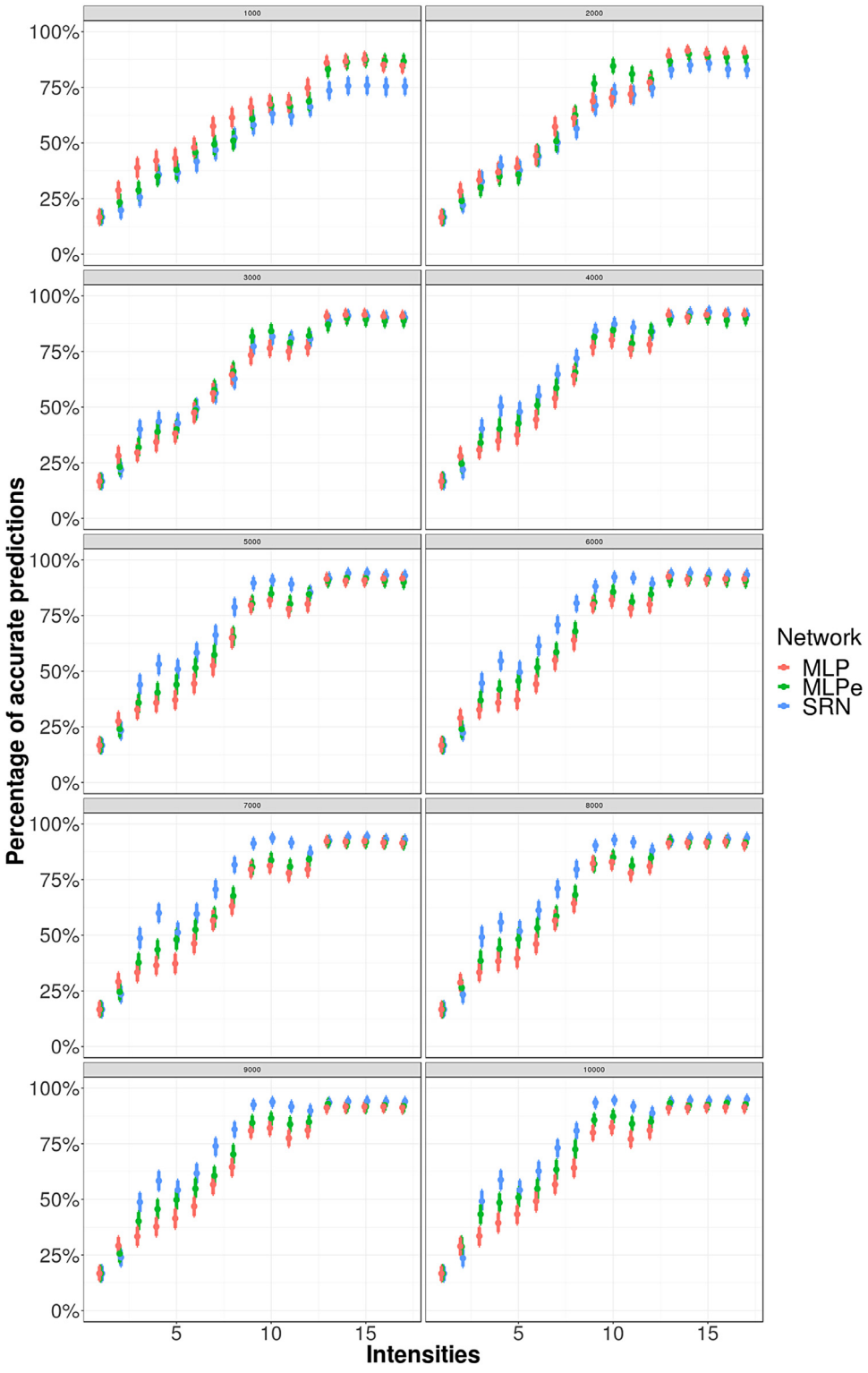

Fig. 9. Mean proportion of correctly anticipated emotions by each type of neural network as a function of the intensity of the emotional image across training epochs. Error bars represent standard errors.

We ran a generalized linear mixed model (GLMM, binomial distribution) to account for the mixed-effects of this study and the non-Gaussian distribution (binary success output) of the data. The model represents the interaction between the neural network type as a between-factor condition and emotions and intensities as within-factors. Primary and interaction effects were explored by means of the Wald Chi-square test, which is an analysis of deviance test more suited to studying binary logistic regression mixed models. Post-hoc analyses, when performed, were calculated according to the least square means method with Holm–Bonferroni correction. Moreover, in order to simplify the statistical analyses, we did not analyze the "Epoch" factor and focused only on the statistics at the end of the training at epoch number 10,000. However, Fig. 9 shows the evolution of the performance through

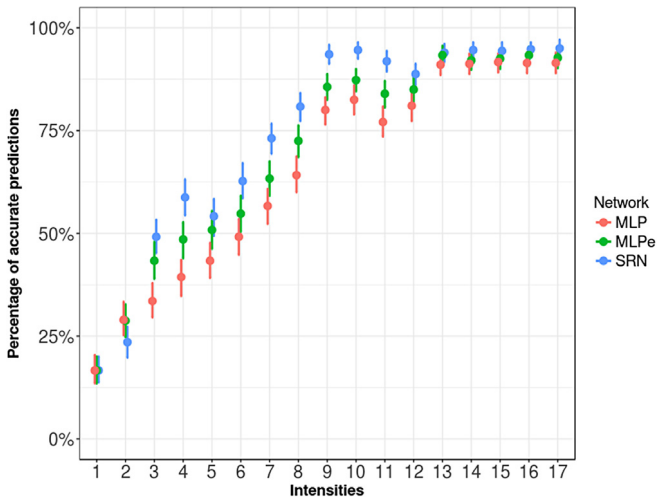

Fig. 10. Mean proportion of correctly anticipated emotions by each type of neural network as a function of the intensity of the emotional image at the end of the training (epoch 10,000). Error bars represent standard errors.

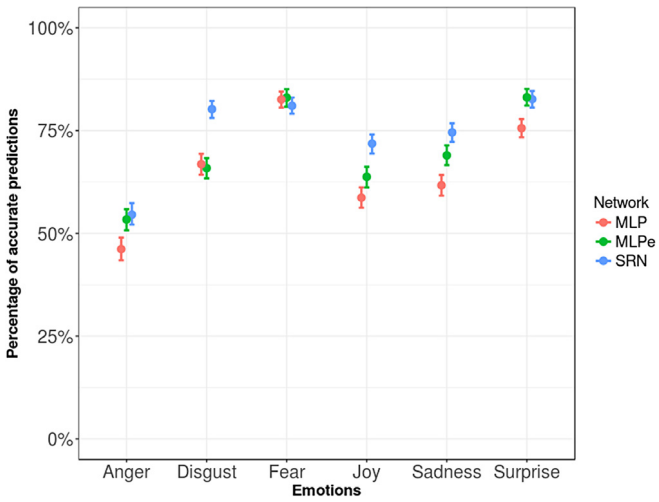

Fig. 11. Mean proportion of correctly anticipated emotions by each type of neural network as a function of the emotional expression. Error bars represent standard errors.

the different epoch numbers. Preliminary analyses including the Epoch factor in GLMM analyses indicated an interaction effect between the neural network type and the epochs (from 1000 to 10,000) that stabilized performance after epoch 7000, with epoch 10,000 reporting the highest average accuracy overall.

The results revealed a main effect of neural network type ($\chi^2(2, N = 240) = 237.02$, $p < 0.0001$) suggesting a better accuracy with SRN compared to MLPe and with MLPe and SRN compared to MLP ($M_{MLP} = 0.65$, $M_{MLPe} = 0.69$, $M_{SRN} = 0.74$). A main effect of intensity ($\chi^2(16, N = 240) = 598.90, p < 0.0001$) was also observed and suggests that prediction performances rise as intensities shift toward the apex (Fig. 10). Moreover, an interaction between neural network types and intensities was also found ($\chi^2(32, N = 240) = 113.62$, $p < 0.0001$), suggesting that the advantage of the SRN over the MLPe and MLP is dependent on the intensity of the facial expression. As expected, the first intensity performances were not significantly different as they are all by chance. SRN accuracy was only lower than the other models in the second intensity. Pairwise comparisons between SRN, MLPe and MLP at each of the remaining intensities indicated an effect in favor of the SRN compared to MLP in all intensities except 13 and 14 (Fig. 10). MLPe had higher accuracy than MLP in all intensities below number 11, apart from intensities 2, 9 and 7.

A main effect of emotion ($\chi^2(5, N = 240) = 192.36, p < 0.0001$) revealed that all emotions were not equally recognized (Fig. 11). The interaction effect between neural network types and emotions (Fig. 11) was significant ($\chi^2(10, N = 240) = 155.90$, $p < 0.0001$). Pairwise comparisons revealed that Fear was predicted at the same accuracy level by all three types of neural networks. SRN significantly outperformed MLP in predicting all other emotions (*ps.* < 0.0001). MLPe performed as well as SRN with regard to Anger and Surprise. SRN outperformed MLPe in the prediction of Disgust, Joy and Sadness (*p.* < 0.0001). MLPe had similar performances to MLP when predicting Disgust.

Quadratic error on predictive performance

In addition to the anticipation accuracy of the neural networks, we investigated the error rate of both types of neural network in predicting the expected values of output neurons at time $t+1$ after exposure to emotional expressions at time t. We ran a Linear Mixed Model (LMM) analysis to account for the mixed-effect of this study. In the same manner, this new model represents the interaction between the neural network types as a between-factor condition

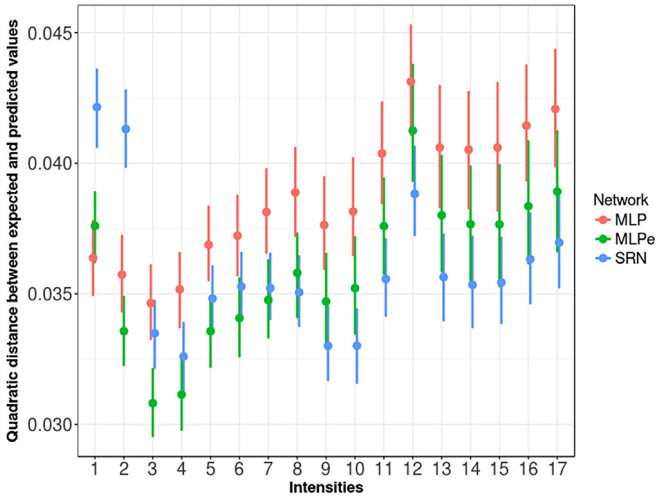

Fig. 12. Mean quadratic error achieved by each type of neural network as a function of the intensity of the emotional image. Error bars represent standard errors.

and the within-factor characteristics of emotions and intensities. Fisher tests were used to retrieve primary and interaction effects, and Holm–Bonferroni corrections were applied for post-hoc comparisons.

Concerning the continuous variable "quadratic error" – defined as the Euclidean distance between output vector and expected vector – results revealed a significant primary effect of neural network type ($F(3,77) = 57.99$; $p < 0.0001$; $M_{MLP} = 0.039$, $M_{MLPe} = 0.036$, $M_{SRN} = 0.036$). Pairwise comparisons revealed higher average performances for MLPe and SRN compared to MLP (*ps* < 0.0001). We observed a significant interaction effect between the neural network types and intensities ($F(48,1309) = 9.49$; $p < 0.0001$), and neural network types and emotions ($F(15,462) = 54.88$; $p < 0.0001$). Post-hoc comparisons between interaction effects and intensities pointed to significantly higher quadratic errors for SRN during the first two intensities (*ps* < 0.0005) and no significant difference between MLP and MLPe. SRN showed errors below MLP for intensity 7 onwards (*ps* < 0.05). Moreover, the SRN error appeared lower but did not statistically differ from that of MLPe (intensity 3 and over) (see Fig. 12).

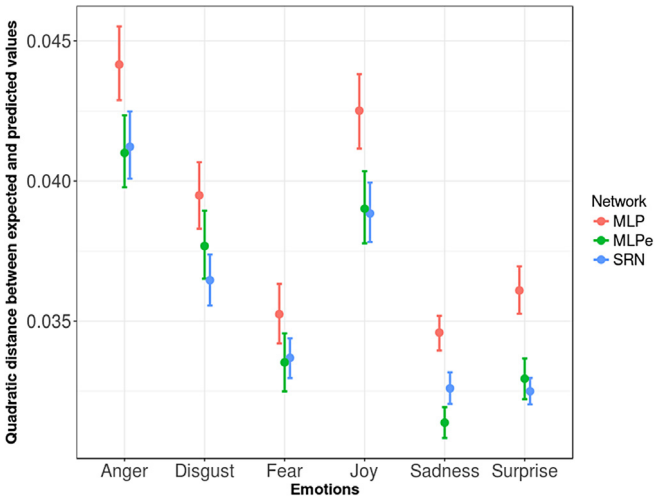

Fig. 13. Mean quadratic error achieved by each type of neural network as a function of the emotional expression. Error bars represent standard errors.

Post-hoc comparisons exploring the differences between interaction effect and emotions showed significantly lower quadratic errors for MLPe and SRN compared to MLP, with regard to all emotions but Fear ($ps < 0.03$). SRN errors were significantly higher than those of MLPe's only in regard to Sadness (see Fig. 13).

4. Discussion

The purpose of the current study was to explore the importance of top-down neural connectivity to (i) categorize dynamic emotional facial expressions, and (ii) anticipate the content of subsequent video frames. In order to evaluate the influence of top-down recurrent connectivity compared to neurobiological human brain data while keeping the maximum degree of comparability between the models, we chose the most simple and comparable types of neural network: a Multi-Layer Perceptron and a Simple Recurrent Network. Thus we avoided more sophisticated, but also technically different, algorithms (i.e. which involve many parameters) such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) networks (Barros, Jirak, Weber, & Wermter, 2015; Sun, Cao, He, & Yu, 2017). Anticipation accuracy was performed by means of a winner-take-all method and anticipation was evaluated by means of a continuous quadratic error.

In terms of anticipation accuracy, the results were not as straight-forward as we would imagine when comparing an SRN to a MLP. The SRN outperformed the MLP as a main effect in this anticipation task, this would be intuitively coherent given the dynamic nature of the inputs. However, we also observed interesting interaction effects with regards to our initial hypotheses. It should be noted that these hypotheses were formulated according to previous neural and psychological data reported in the scientific literature. First, the interaction between the type of neural network and the intensity of the emotional expression revealed that the SRN does not surpass the MLP for the intense expressions. Pairwise comparisons between the SRN and the MLP clearly indicated an advantage for the intermediate intensities of emotional facial expressions (in other words, the most subtle expressions). What complements this finding in Simulation 1 is that the interaction between the type of neural network and the content of emotional expressions revealed that the advantage of the SRN over the MLP exists for ambiguous emotional expressions (fear, anger, etc.) but not for the easier emotional expressions (e.g. happiness). This finding is congruent with the fact that happiness is better recognized at a computational level for artificial neural networks (Dailey et al., 2002; Mermillod, Bonin et al., 2010; Mermillod, Droit-Volet et al., 2010; Mermillod, Vermeulen et al., 2009), but also at a psychological and physiological level for humans (Beffara et al., 2012; Mermillod et al., 2011, 2017). However, given that this advantage of the SRN over the MLP and MLPe for ambiguous emotions (compared to happiness) was not obtained in Simulation 2, we cannot ensure that this advantage does not actually depend on the database. This finding has to be confirmed with a larger database including more ambiguous emotions than these six basic emotional expressions.

Regarding the errors produced by the neural networks during the prediction of the next frame of the video, we observed that the SRN outperformed the MLP, as indicated by the main effect of the neural network architecture. However, interestingly, and as found for accuracy, this effect on the continuous error variable was related to the specific intensity of an emotional expression, mainly intermediate intensities, when the signal was sufficient to detect and anticipate an emotional expression but still too subtle to be efficiently processed by a pure bottom-up neural network. As shown in Fig. 4, the performances of the MLP and SRN showed similar results for the last intensity of the emotional expression, when the intensity of the EFE was very strong.

Moreover, concerning the general superiority of the SRN compared to the MLP in its capacity to predict the next frame of emotional expressions (assessed by the average quadratic distance between expected and observed output, Fig. 7), we observed a significant difference toward less errors produced by the SRN compared to the MLP for anger, fear and happy facial expressions. Other differences, albeit in favor of the SRN, were not significant. In other words, although the SRN was not superior to the MLP for happiness in terms of accuracy (we assume this to be due to a ceiling effect of both MLP and SRN which almost perfectly recognized happiness with this database), results in terms of prediction performance revealed that the SRN performed better than the MLP, especially in cases of subtle or ambiguous emotional expressions (Dailey et al., 2002; Mermillod et al., 2011). In other words, the more difficult it is to anticipate the emotional expression, the more important the use of recurrent neural connections to improve the performance of the neural network. Therefore, we can assume that this top-down neural connectivity, recently discovered for humans (Bar et al., 2006; Kauffmann, Bourgin et al., 2015; Kauffmann, Chauvin et al., 2015), could have been very important in the history of human evolution for the efficient recognition and anticipation of emotions expressed by congeners. Moreover, we can assume that the importance of this feed-back connectivity could be a simple example of a more general advantage of those recurrent processes for different cognitive tasks (e.g. anticipation and categorization of various types of perceptual or emotional events). In accordance with our initial hypotheses, it appears that the top-down connectivity from the associative to the visual categorization areas that are observed in the human brain (Bar, 2004; Kauffmann, Bourgin et al., 2015; Kauffmann, Chauvin et al., 2015) could be an efficient way to improve the anticipation accuracy and the anticipation of emotional events such as dynamic emotional facial expressions. Complementary, subsequent lines of research will need to determine if the spatial frequency characteristics of these top-down connections, provided by fast-track neural pathways to the orbitofrontal cortex, on the basis of low spatial frequency information as top-down signal, can improve the efficiency of an artificial neural network. Neurocomputational evidence points to a preferential advantage for the use of LSF information during the categorization of emotional facial expressions by artificial neural networks (Cipollini & Cottrell, 2014; Mermillod, Bonin et al., 2010; Mermillod, Vuilleumier et al., 2009; Wang & Cottrell, 2013). It

remains to be determined if LSF is more efficient than high spatial frequencies for predicting upcoming signals based on top-down recurrent processes, as reported in the current article. Our results also fall well within the predictive coding framework (Friston, 2005; Rao and Ballard, 1999) which posits that perception relies on a permanent comparison between sensory signals and elaborated predictions based on learnt regularities in the environment and prior experiences. According to this theoretical framework, such a mechanism would be particularly relevant when sensory inputs are ambiguous or noisy since predictions could be used to constrain processing and interpretation (Summerfield & Egner, 2009). The results of the present study support that view by showing that the use of prior information mostly facilitated the anticipation of ambiguous stimuli.

Another important line of research will be to determine to what extent the importance of recurrent synaptic connections in the anticipation of emotional facial expressions could be generalized for dynamic stimuli for different scales, orientations and perceptual properties. This question should also be addressed when dangerous stimuli are provided (e.g. a predator or a moving vehicle) rather than emotional expressions. This could be achieved by means of more sophisticated neural networks such as DNN or CNN. While the current scientific literature on simple recurrent networks is sufficient to explain the efficiency of recurrent connections on the anticipation of dynamic events (Iwasaki & Furukawa, 2016; Zhang, Wang, & Liu, 2007, 2008, 2014), "static" (e.g. DNN) and "recurrent" (e.g. LSTM) deep learning models could be more difficult to compare and understand given the large number of layers and parameters. This has to be tested on more general stimuli in further studies.

Acknowledgments

We thank Roisin Lee and Richard May for the careful reading of the English on the preprint of this article. Louise Kauffmann was supported by NeuroCoG IDEX UGA in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02).

References

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289. http://dx.doi.org/10.1016/j.tics.2007.05.005.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., et al. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454.

Barrett, L. F., & Bar, M. (2009). See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences*, *364*(1521), 1325–1334.

Barros, P., Jirak, D., Weber, C., & Wermter, S. (2015). Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks*, *72*, 140–151.

Beffara, B., Ouellet, M., Vermeulen, N., Basu, A., Morisseau, T., & Mermillod, M. (2012). Enhanced embodied response following ambiguous emotional processing. *Cognitive Processing*, *13*(1), 103–106.

Beffara, B., Wicker, B., Vermeulen, N., Ouellet, M., Bret, A., Funes, M. J., et al. (2015). Reduction of interference effect by low spatial frequency information priming in an emotional stroop task. *Journal of Vision*, *15*(6), 16 1-9.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, *3*, 1137–1155.

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, *36*(2), 96–107.

Cipollini, B., & Cottrell, G. (2014). A developmental model of hemispheric asymmetry of spatial frequencies. In *Proceedings of the annual meeting of the cognitive science society*, *Vol. 36*, No. 36.

Dailey, M. N., & Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks*, *12*(7–8), 1053–1074.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience*, *14*(8), 1158–1173.

Ekman, P., & Friesen, W. (1976). *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: Investigatoris guide*. Consulting Psychologists Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B Biological Sciences*, *360*(1456), 815–836. http://dx.doi.org/10.1098/rstb.2005.1622.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

Iwasaki, T., & Furukawa, T. (2016). Tensor SOM and tensor GTM: Nonlinear tensor analysis by topographic mappings. *Neural Networks*, *77*, 107–125.

Jones, J. P., & Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*, 1187–1211.

Jones, J. P., Stepnoski, A., & Palmer, L. A. (1987). The two-dimensional spectral structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1212–1232.

Kauffmann, L., Bourgin, J., Guyader, N., & Peyrin, C. (2015). The neural bases of the semantic interference of spatial frequency-based information in scenes. *Journal of Cognitive Neuroscience*, 1–10.

Kauffmann, L., Chauvin, A., Pichat, C., & Peyrin, C. (2015). Effective connectivity in the neural network underlying coarse-to-fine categorization of visual scenes. A dynamic causal modeling study. *Brain and Cognition*, *99*, 46–56.

Kauffmann, L., Ramanoël, S., & Peyrin, C. (2014). The neural bases of spatial frequency processing during scene perception. *Frontiers in Integrative Neuroscience*, *8*.

Kawasaki, H., Adolphs, R., Kaufman, O., Damasio, H., Damasio, A. R., Granner, M., et al. (2001). Single-neuron responses to emotional visual stimuli recorded in human ventral prefrontal cortex. *Nature Neuroscience*, *4*(1), 15–16.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kveraga, K., Boshyan, J., & Bar, M. (2007). Magnocellular projections as the trigger of top-down facilitation in recognition. *Journal of Neuroscience*, *27*(48), 13232–13240.

Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain and Cognition*, *65*(2), 145–168.

Li, H., Lin, Z., Shen, X., Brandt, J., & Hua, G. (2015). A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5325–5334).

Mermillod, M., Auxiette, C., Chambres, P., Mondillon, L., Galland, F., Jalenques, I., et al. (2011). Contraintes perceptives et temporelles dans l'exploration du modèle de ledoux. *L'année Psychologique*, *111*(3), 465–479.

Mermillod, M., Bonin, P., Mondillon, L., Alleysson, D., & Vermeulen, N. (2010). Coarse scales are sufficient for efficient categorization of emotional facial expressions: Evidence from neural computation. *Neurocomputing*, *73*, 2522–2531.

Mermillod, M., Droit-Volet, S., Devaux, D., Schaefer, A., & Vermeulen, N. (2010). Are coarse scales sufficient for fast detection of visual threat?. *Psychological Science*, *21*(10), 1429–1437.

Mermillod, M., Grynberg, D., Lopez, L. Pio., Rychlowska, M., Beffara, B., Vermeulen, N., et al. (2017). Evidence of rapid modulation by social information of subjective, physiological and neural responses to emotional expressions. *Frontiers in Behavioral Neuroscience*, *11*, 231.

Mermillod, M., Vermeulen, N., Lundqvist, D., & Niedenthal, P. M. (2009). Neural computation as a tool to differentiate perceptual from emotional processes: The case of anger superiority effect. *Cognition*, *110*(3), 346–357.

Mermillod, M., Vuilleumier, P., Peyrin, C., Alleysson, D., & Marendaz, C. (2009). The importance of low spatial frequency information for recognizing fearful facial expressions. *Connection Science*, *21*(1), 75–83.

Niedenthal, P. M., Mermillod, M., Maringer, M., & Hess, U. (2010). The simulation of smiles (SIMS) model: Embodied simulation and the meaning of facial expression. *Behavioral and Brain Sciences*, *33*(06), 417–433.

O'Callaghan, C., Kveraga, K., Shine, J. M., Adams, R. B., & Bar, M. (2017). Predictions penetrate perception: Converging insights from brain, behaviour and disorder. *Consciousness and Cognition*, *47*, 63–74.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015, September). Deep face recognition. In *BMVC*, *Vol. 1*, No. 3, (p. 6).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., & DeVito, Z., et al. (2017). Automatic differentiation in pytorch. In *NIPS-W*.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. http://dx.doi.org/10.1038/4580.

Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1985). Learning internal representations by error propagation (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Sherman, S. M., & Guillery, R. W. (2002). The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences*, *357*(1428), 1695–1708.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409.

Sun, B., Cao, S., He, J., & Yu, L. (2017). Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy. *Neural Networks*. http://dx.doi.org/10.1016/j.neunet.2017.11.021.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014, June). Deepface: Closing the gap to human-level performance in face verification. In *Computer vision and pattern recognition (CVPR), 2014 IEEE conference* (pp. 1701–1708).

Tong, M. H., Joyce, C. A., & Cottrell, G. W. (2008). Why is the fusiform face area recruited for novel categories of expertise? A neurocomputational investigation. *Brain Research*, *1202*, 14–24.

Trapp, S., & Bar, M. (2015). Prediction, context, and competition in visual recognition. *Annals of the New York Academy of Sciences*, *1339*(1), 190–198.

Van Der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving faces, looking places: validation of the Amsterdam dynamic facial expression set (ADFES). *Emotion*, *11*(4), 907.

Wang, P., & Cottrell, G. (2013). A computational model of the development of hemispheric asymmetry of face processing. In *Proceedings of the annual meeting of the cognitive science society, Vol. 35*, No. 35.

Wang, P., & Cottrell, G. W. (2016). Basic level categorization facilitates visual object recognition. In *4th International conference on learning representations (ICLR 2016) workshop* arXiv:1511.04103.

Wang, P., & Cottrell, G. W. (2017). Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. *Journal of Vision*, *17*(4) 9–9.

Wang, P., Gauthier, I., & Cottrell, G. (2016). Are face and object recognition independent? A neurocomputational modeling exploration. *Journal of Cognitive Neuroscience*, *28*(4), 558–574.

Zhang, H., Wang, Z., & Liu, D. (2007). Robust exponential stability of recurrent neural networks with multiple time-varying delays. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *54*(8), 730–734.

Zhang, H., Wang, Z., & Liu, D. (2008). Global asymptotic stability of recurrent neural networks with multiple time-varying delays. *IEEE Transactions on Neural Networks*, *19*(5), 855–873.

Zhang, H., Wang, Z., & Liu, D. (2014). A comprehensive review of stability analysis of continuous-time recurrent neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(7), 1229–1262.