



Title	Convolutional Neural Network Models for Subcortical Processing of Facial Expression
Author(s)	Lim, Chanseok
Citation	大阪大学, 2023, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/92999
rights	Reproduced with permission from Springer Nature
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

Doctoral Thesis

**Convolutional Neural Network Models
for Subcortical Processing of Facial Expression**

皮質下顔表情処理のための
畳込みニューラルネットワークモデル

Chanseok Lim

Graduate School of Frontier Biosciences

Osaka university

September 2023

Table of Contents

Summary	4
Introduction.....	6
1.1 Two facial perceptions	7
1.2 Computational approaches for visual processing.....	12
1.3 Aim of the study.....	15
Materials and Methods.....	18
2.1 Architecture of SNNs.....	18
2.2 Face images and training of the SNNs.....	24
2.3 Test for reference frames of SF tuning of model units	27
2.4. Analysis of the effects of the max pooling operation on SF selectivity.....	29
Results.....	31
3.1 Performance of SNNs in facial expression classification	31
3.2 Effects of modification of SNNs on classification performance	34
3.3 Spatial frequency representation in the FC1 layer	37
3.4 Training and testing of SNNs with a different face database.....	44
3.5 Effects of max pooling on SF tuning	47
3.6 Mathematical formulation of the max pooling operation	50
3.7 Ablation of max pooling	55
3.8 Effects of sliding stride on SF tuning.....	57
Discussion	60
4.1 Modest performance of the SNNs and roles of the subcortical pathway properties.....	61
4.2 Confusions of facial expressions in the SNNs, DNNs and patients.....	63

4.3 Reference frame of coding SF information.....	65
Conclusion	69
Appendix.....	71
<i>A1 SF tunings of DoG filters</i>	71
Acknowledgement	74
References.....	75
Personal Information.....	86
Bibliography.....	87

Summary

Perception of facial expression is crucial in the social life of primates. This visual information is processed along the ventral cortical pathway and the subcortical pathway. They process information in parallel, and finally meet at the amygdala. The ventral cortical pathway consists of a network of areas in the occipito-temporal region of the cerebral cortex (e.g., visual areas V1, V2, V4, and inferior temporal cortex, IT), and its processing of face information is slow but precise. The subcortical pathway is composed of a few processing stages of phylogenetically ancient regions; the superior colliculus in the midbrain, the pulvinar nucleus of the posterior thalamus, and the amygdala in the medial limbic system. The subcortical processing is fast but coarse. Though the difference of processing speeds is explained by the difference in the number of processing stages between the two pathways, it is unclear whether the difference in the number of stages also leads to the difference of processing granularities. What computational properties in the subcortical pathway make its processing coarse-grained? What computational models faithfully mimic the subcortical processing?

To address these questions, I constructed convolutional neural networks incorporating three prominent properties of the subcortical pathway; shallow layer architecture, concentric receptive fields at the first stage, and a greater degree of spatial pooling. I trained these networks, referred to as shallow neural networks (SNNs), and their modified versions to classify seven facial expressions (angry, disgusted, fearful, happy, sad, surprised, and neutral), analyzed their performance, and examined the internal representation of spatial frequency (SF) information across computational units of the final processing layer.

The SNNs can be trained to classify the seven facial expressions with a correct rate of 51% (chance level, 14%). The performance was well above chance, but substantially below perfect. This modest performance was gradually improved by replacing the three properties, one-by-one, two together, or all three together, with the corresponding features in the cortical pathway; additional convolution layer, Gabor-type filters at the first convolution layer, and narrower pooling windows. The results indicate that all three subcortical features are essential for the coarse processing. The effects of the three features on the classification performance were partially additive, suggesting that the three features exerted their effects partially independently.

A previous study (Inagaki and Fujita, 2011) revealed a prominent difference in the reference frame of neuronal tuning to SFs between the two pathways. Neurons in the IT, the final stage of

the ventral cortical pathway, are tuned to object-based SFs (cycles/object) and represent face patterns in a size-invariant, hence distance-invariant, manner. In contrast, responses of a major population of amygdala neurons are affected by retina-based SFs (cycles/degree). Some units in the final layer of the SNNs were sensitive to SFs in the retina-based reference frame, whereas others were sensitive to object-based SF, in a similar way to neurons in the amygdala. Replacement of any one of the three properties changed the reference frames of units in the final layer. The modified models with added layers or Gabor-type filters reduced the units with the retina-based SF selectivity. On the other hand, units in the models with the shallowness, DoG-type filters, and the narrow pooling showed various selectivities between the two reference frames. The results suggest that both shallow architecture and DoG-type filters were necessary for preserving the sensitivity of the final layer units to the retina-based SFs, and a greater pooling window hindered creation of the intermediate representation between the two reference frames.

In the SNNs and the narrow-pooling models, units tuned to low SFs encoded object-based SFs, whereas units tuned to high SFs encoded the retina-based SFs. Reasoning that non-linear operation is a cause of these effects on the unit responses, I developed a formal description of max pooling. My mathematical analysis revealed that the shift invariance by the max pooling operation led to the size and homogenous invariance. I verified this effect with experimental analysis; outputs from the pooling layer were more similar between the different sizes (i.e., became size-invariant) than the inputs, only when the inputs were composed of low SFs.

I provided the first computational model for facial expression processing in the subcortical pathway. Despite the celebrated success of deep neural networks (DNNs) in modeling visual processing in the ventral cortical pathway, it has remained unclear whether and how the convolutional neural network architecture can be adapted to processing in the subcortical pathway. I demonstrated that the SNNs implemented with the three computational properties of the subcortical pathway successfully learn facial discrimination with a modest correct rate. The three properties are all essential for reproducing the modest performance by V1-lesioned patients who discriminate facial expressions with the subcortical pathway. These properties also necessary for reproducing the representation of SFs in the retina-based coordinate observed in a population of amygdala neurons. Research interests on the role of the subcortical structures in cognitive functions has recently surged, but physiological data are still much sparser for the subcortical structures than for the cerebral cortex. Computational approach such as the one I present here is expected to complement the sparseness of the data and guide future research.

Introduction

Most animals engage in interactions with other beings of their own and other species, forming various relationships such as predator and prey, leader and follower, parent and offspring, friends, lovers, neighbors, and so on. These interactions often result in the formation of groups known as societies. Communication is a crucial aspect of social life, as it allows individuals to survive and thrive. There are various forms of communication, including vocalization, touch, gesture, and facial expression, which are typically achieved through the transmission of physical signals that are received by sensory organs and perceived by the brain. How to sense and how to perceive have been one of main interests of biology and related fields. In neuroscience, researchers have endeavored to understand the nervous systems underpinning senses and perceptions. visual perception has attracted researcher because our visual information is certainly rich and seems to contain crucial information. Among various aspects of perception, facial perception holds a special place. Certain neurons in the inferior temporal (IT) cortex, referred to as "face cells," exhibit selective activation in response to facial visual stimuli (Taubert et al., 2020). Furthermore, it is remarkable that infants, despite having immature cortex, already show a preference for directing their gazes towards faces within as little as 72 hours after birth (Cassia et al., 2001), highlighting the early significance of facial perception in human development."

To reveal how the brain work for facial perceptions, neuroscientists have employed a variety of approaches, including anatomical, histological, electrophysiological, psychophysical methods. In recent years, computational approaches have gained increasing prominence, following recent celebrated successes of Convolutional Neural Networks. These approaches seek to describe complex neural mechanisms using simplified models, which can provide not only explanations

but also insights for neural mechanisms. Works for cortical face processing and attempts to model the processing, indeed, have significantly advanced our understanding of the cortical processing. However, much of the subcortical processing remain poorly understood, although it would play important role for developments of our visual capabilities and underpin our unconscious behaviors on facial perceptions. In this dissertation, developing computational models, I aimed to understand the subcortical face processing and shed light on potential avenues for future research.

1.1 Two facial perceptions

There exist two modes of facial perceptions: fast and slow ones. In daily lives, we can effortlessly perceive other's faces, and this perception is enough detailed to provide us with inference about other's emotional states from their facial expressions. Facial perception includes obviously includes two different parts: detection and recognition. Face detection must precede recognition because positional information of face is necessary for the detailed analysis. Further, previous research in neuroscience indicates the fast and slow perception are underpinned by different neural systems and the fast one works beyond just detection as rough recognition.

In primates, including humans, two neural pathways, the ventral cortical pathway and the subcortical pathway, contribute to this perceptual ability (Fig. 1.1.1; Pessoa and Adolphs, 2010; Tamietto and de Gelder, 2010; Petray and Bickford, 2019). The ventral cortical pathway consists of a network of areas in the occipito-temporal region of the cerebral cortex, and processes a variety of visual features of objects, people, and environments, including shape, color, texture, material properties, and binocular disparity (Ungerleider and Mishkin, 1982; Connor et al., 2007; Conway et al., 2010; Roe et al., 2012; Kravitz et al., 2013; Vaziri et al., 2014; Verhoff et al., 2016; Komatsu and Goda, 2018). Neurons that preferentially respond to face images are found in several clusters

along this pathway (Haxby et al., 2000; Tsao and Livingstone, 2008; Duchaine and Yovel, 2015; Freiwald et al., 2016). They constitute the neural system that analyzes facial details such as expression, identity, and direction of attention. The subcortical pathway consists of a few processing stages of phylogenetically ancient regions: the superior colliculus in the midbrain, the pulvinar nucleus in the posterior thalamus, and the amygdala in the medial limbic system. The subcortical pathway is suggested to mediate rapid behavioral and physiological (autonomic) responses to sensory signals related to possible dangers such as fearful faces (Tamietto and de Gelder, 2010; Nakano et al., 2013; for a critical review, see Pessoa and Adolphs, 2010). The ventral cortical pathway and the subcortical pathway intersect at the amygdala.

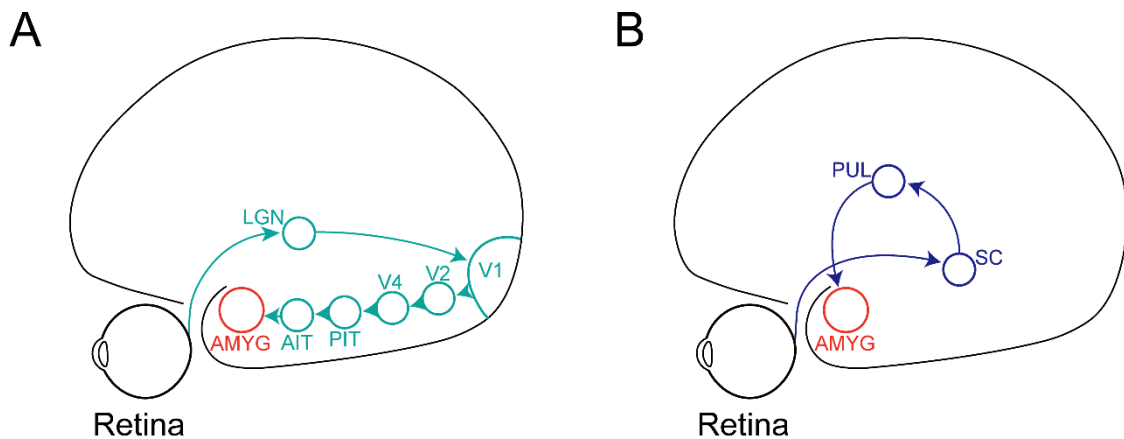


Figure 1.1.1. Two visual pathways in primate's brains. A. The geniculostriate pathway and the ventral cortical pathway. Light signals are transformed into electric signals by retina and transmitted to the primary visual cortex (V1) via a relay in the lateral geniculate nucleus (LGN) of the thalamus. The ventral cortical pathway consists of a network of areas in the occipito-temporal region of the cerebral cortex. This pathway carries information from V1 to the amygdala in the medial limbic system (AMG) through cortical regions such as V2, V4, and higher visual area PIT and AIT. B. The subcortical pathway. The subcortical pathway consists of a few

processing stages of phylogenetically ancient regions: the superior colliculus in the midbrain (SC), the pulvinar nucleus in the posterior thalamus (PUL), and the amygdala.

Facial perceptions of the ventral cortical pathway.

Face has very complex visual property. Face consists of several essential parts, including eyes, nose, mouth, and ears, and those parts must be arranged on predetermined positions. face-responsive neurons in IT, which is often referred as “face cells” deliver judgements based on these complex conditions. Their selectivity is constructed by gradual conversion of visual properties along the ventral cortical pathway. The ventral cortical pathway begins in the primary visual cortex (V1) in the occipital lobe. Neurons of V1 receive visual information as electric signals, which are transformations of light signals originated in the external world. The light signals are first received by photoreceptor cells in the retina and transmitted to V1 via a relay in the lateral geniculate nucleus (LGN) of the thalamus. V1 neurons have localized receptive fields and detects specific visual properties, such as orientation, spatial frequency (SF), color, and binocular disparity. These localized receptive fields are constructed by collecting excitatory and/or inhibitory signals from LGN neurons existing in the spatial neighborhood each other. In other words, V1 neurons codes visual information about inputted positions and specific visual properties. This information, next, transmitted to V2 and V2 neurons greatten the localization of their RFs and transform visual properties into more complex ones. In this way, simple visual properties information coded on V1 are converted into complex properties on IT, such as face, along the ventral cortical pathway.

Whereas face-responsive neurons in IT have a various kind of selectivities about faces, including facial expression and face identity (Hasselmo et al., 1989; Sugase et al., 1999), one of notable

property of non-selectivities is size invariance. Face-responsive neurons are poorly affected by stimulus sizes on the retina (Rolls and Baylis, 1986), although different sizes of a face provide different retinal images. This invariance is explained with tuning properties of face-responsive neurons to object-based SFs (Fig. 1.1.2.; Inagaki and Fujita, 2011). Object-based SF (cycles/object) is one of reference frames for SF and quantifies cyclic luminance patterns of lights and darks relatively to object sizes, such as face sizes in stimulus. This relative SFs are obtained by normalization of absolute retina-based SFs (cycles/degree). Retina-based SF is quantification of cyclic luminance patterns to visual angles. Previous studies in psychophysics demonstrated performance of face recognitions is more affected by object-based SF than retina-based (Hayes et al., 1986; Näsänen, 1999; for a review, see Sowden and Schyns, 2006). Afraz and his/her colleagues demonstrated a more casual relationship between activities of face-responsive neurons and face perceptions (Afrac et al., 2006). They found artificial activations of face-responsive neurons increased likelihood that macaque monkeys classified ambiguous noise stimulus as a face (also see Sadagopan et al., 2017).

These information about faces is carried from IT to the amygdala. There exist anatomical inputs from the anterior inferotemporal area (Amaral and Price, 1984; Cheng et al., 1997) and face-responsive neurons are also found in amygdala (Leonard et al., 1985; Nakamura et al., 1992, Gothard et al., 2007; Rutishauser et al, 2011). In addition, Inagaki and his/her colleagues showed decoding performances of facial expressions from population activities of amygdala neurons are nearly same with those from population activities of IT neurons at around 160 ms from stimulus onset (Inagaki et al., 2022a).

Facial processing of the subcortical pathway.

The amygdala receives inputs not only from cortical but also subcortical pathways and the subcortical one also carries facial information. The subcortical pathway serves as rapid responses to information related with danger. This idea originated from a generalization of findings in fear-conditioning studies in rats by LeDoux (LeDoux, 2000).

In the subcortical pathway, retinal signals immediately transmitted to the amygdala through relays on superior colliculus and plvinar (Elorette et al., 2018; Inagaki et al., 2022a). Psychological and brain imaging studies suggest that the subcortical pathway subserves the ability of some patients with lesions in the primary visual cortex (V1) to discriminate facial expressions despite lacking visual awareness (“affective blindness”; de Gelder et al., 1999; Pegna et al., 2005; Striemer et al., 2019). These patients also automatically exhibit specific facial expressions and pupillary reactions when exposed to fearful or happy faces (Tamietto et al., 2009). Studies have also shown that the subcortical pathway supports unconscious face perception in neurologically healthy subjects (Morris et al., 1999, 2001, Whalen et al., 1998). Furthermore, orientation bias toward faces or face-like patterns by newborn babies is suggested to be mediated by the subcortical pathway (Cassia et al., 2001; Johnson, 2005; but see Buiatti et al., 2019). Importantly, these perceptual abilities are not perfectly accurate, instead resulting in modest performance at above-chance levels. These findings suggest that information on faces conveyed by the subcortical pathway is coarser than that carried by the ventral cortical pathway.

Electrophysiological studies have demonstrated that processing of facial expression in the subcortical pathway is indeed fast and coarse. Méndez-Bértolo and colleagues (2016) showed that intracranial local field potentials in the human amygdala respond differentially to fearful faces versus other faces within 74 ms after stimulus onset. A recent single-neuron recording study in

the monkey revealed that a population of amygdala neurons responded to threatening faces even within 50 ms (Inagaki et al., 2022a). This early response, when combined across an ensemble of neurons, carries information that allows linear classifiers to discriminate threatening faces from neutral and affiliative faces. The rate of correctly discriminating the three expressions is around 50%; this is well above chance (33%), but significantly worse than perfect. Some neurons in the superior colliculus and pulvinar also respond to faces and face-like patterns with a short latency of 30-50 ms (Nguyen et al., 2013, 2014).

1.2 Computational approaches for visual processing

As mentioned earlier, the two visual pathways in the brain consist of distinct neural systems and carry different types of facial information. To gain further insights into the functional aspects of these systems, one possible approach is to develop computational models that simulate their behavior. Computational models aim to represent biological neural systems using basic operations that can be implemented on computers. One type of computational model commonly used for simulating biological neural systems is artificial neural networks (ANNs). ANNs are comprised of artificial neurons, also known as units, which perform calculations by summing weighted inputs and applying activation functions to the resulting sum. They then output the values obtained from the activation functions. Convolutional neural networks (CNNs) are a specific class of artificial neural networks that were originally inspired by the structure and functionality of neural systems.

CNNs have proven to be not only powerful image processing systems but also successful computational models for brain visual processing. The origin of CNNs dates back to 1980 when Kunihiko Fukushima developed the first system known as "neocognitron". Inspired by the

functions of simple cells and complex cells in the primary visual cortex (V1), this system was the first to implement convolution layers and pooling layers (Fukushima, 1980). Convolution layers perform the detection of localized visual properties using filters, which are mini-images that encode specific visual features. These filters mimic the receptive fields of V1 simple cells. Pooling layers, on the other hand, conduct down-sampling similar to V1 complex cells.

CNNs learn to classify inputs by adjusting the connection weights between computational units through supervised learning (LeCun et al., 2015). Typical CNNs, known as deep neural networks (DNNs), consist of several to tens of layers. The power of DNNs was demonstrated in the ImageNet Large Scale Visual Recognition Challenge 2012, where Krizhevsky and colleagues won the prize with a DNN called AlexNet achieving approximately 10% lower error rate than the previous year's challenge winner. Furthermore, it is remarkable that AlexNet learned Gabor-like filters at its first convolution layer solely through category classification learning, without any prior domain knowledge. This finding had a significant impact on neuroscience, suggesting that developing similar architectures and optimizing them for specific tasks could lead to artificial systems that resemble brain-like systems. Furthermore, it implied that our own visual systems may have evolved through a process of optimization for specific visual tasks. Since then, the interaction between neuroscience and computer vision has been vibrant, with CNNs providing valuable insights into visual processing in the ventral cortical pathway (Yamins and DiCarlo, 2016; Hassabis et al., 2017).

For instance, CNNs with biologically plausible architectures have been shown to acquire similarities in inner information representations not only between the top layer and the inferotemporal cortex (IT) but also between middle layers and area V4 (Yamins et al., 2014).

CNNs have also revealed that visual properties encoded at different stages in the ventral cortical pathway gradually increase in complexity along the pathway (Güçlü and van Gerven, 2015). Furthermore, Kar and colleagues found that challenging images for DNNs also pose challenges for monkeys, and the superior performance of monkeys in classifying difficult images can be attributed to the sustained firing rates of IT neurons (Kar et al., 2019).

These celebrated successes of modeling the ventral cortical pathway with CNNs allowed for the exploration of questions about the subcortical processing of facial expressions. Specifically, what architectural and physiological properties of the subcortical pathway are responsible for its fast, coarse processing? The fast processing most likely arises from the small number, or “shallowness,” of processing stages in the subcortical pathway, given that the ventral cortical pathway and its upstream area (the lateral geniculate nucleus) consist of a larger number of regions (at least six before reaching the amygdala) than the subcortical pathway (only two), and that every transition from one region to the next takes at least 10 ms (Schmolesky et al., 1998). It is unclear whether the shallow processing similarly explains the coarseness of the information transmitted by the subcortical pathway to the amygdala. This uncertainty arises from the fact that in addition to the difference in the number of processing stages, visual response properties differ markedly between the two pathways.

The subcortical pathway has differences in not only its number of processing stages from that of the ventral cortical pathway. Receptive field organizations and its sizes are also different. Neurons in the superior colliculus at the first stage of the subcortical pathway show circular receptive fields with center-surround organization, which can be modeled using the difference-of-Gaussian (DoG) function (Cyndager and Berman 1972; Updyke 1974; Marino et al. 2008; Churan et al.,

2012). By contrast, simple cells of V1 at the first stage of the cortical pathway have elongated receptive fields with side-by-side ON and OFF subregions, which can be modeled by two-dimensional Gabor functions (Jones and Palmer, 1987). Furthermore, the receptive field is typically larger in the superior colliculus (for the foveal field, $1.5\text{--}10^\circ$ in superficial layers, $10\text{--}20^\circ$ in deep layers; Goldberg and Wurtz, 1972; Wallace et al., 1997) than in V1 (1.3° in complex cells; Van den Bergh et al., 2010) and the extrastriate areas V2 and V4 (Freeman and Simoncelli, 2011). Thus, spatial pooling across ascending stages occurs over a wider visual field area in the subcortical pathway than in the ventral cortical pathway.

1.3 Aim of the study

In the present study, I addressed how three properties of the subcortical pathway, i.e., the shallowness of processing stages, DoG-type receptive fields at the initial stage, and spatial pooling over a wider visual field, influence facial expression processing. To this aim, I constructed convolutional neural networks (CNNs) and analyzed their performance in facial expression discrimination. I designed my CNNs to imitate the subcortical pathway by reducing the number of processing stages and by implementing DoG-type receptive fields and a wider extent of pooling. These CNNs, hereafter referred to as shallow neural networks (SNNs), learned to discriminate facial expressions with modest correct rates. Replacing the three properties, one-by-one, two at a time, or all three simultaneously, with the corresponding properties in the ventral cortical pathway improved discrimination performance, suggesting that all three features are necessary to explain the coarseness of face processing in the SNNs. I further showed that like some neurons in the amygdala, a major group of units in the final processing layer of the SNNs were sensitive to spatial frequency (SF) in the retina-based reference frame. Mathematical formulation and simulation revealed that pooling across a wider area is a key process underlying this property.

Object-based vs. retina-based special frequency representation.

A prominent difference in visual responses between the two pathways is the reference frame of neuronal tuning to SFs (Inagaki and Fujita, 2011). Neurons in IT, the final stage of the ventral cortical pathway, are tuned to object-based SFs (cycles/object) and represent face patterns in a size-invariant, hence distance-invariant, manner (Fig. 1.1.2B, left). Thus, the ventral cortical pathway converts the representation of SFs in the retinal coordinate (cycles/degree) to that of object-based SFs. By contrast, many amygdala neurons preserve sensitivity to retina-based SFs and are tuned to a range of them. When the stimulus size is changed, they change their preferred object-based SFs (Fig. 1.1.2B, right). I analyzed the reference frame of FC1 SF tuning to evaluate how well my models captured this characteristic of subcortical processing.

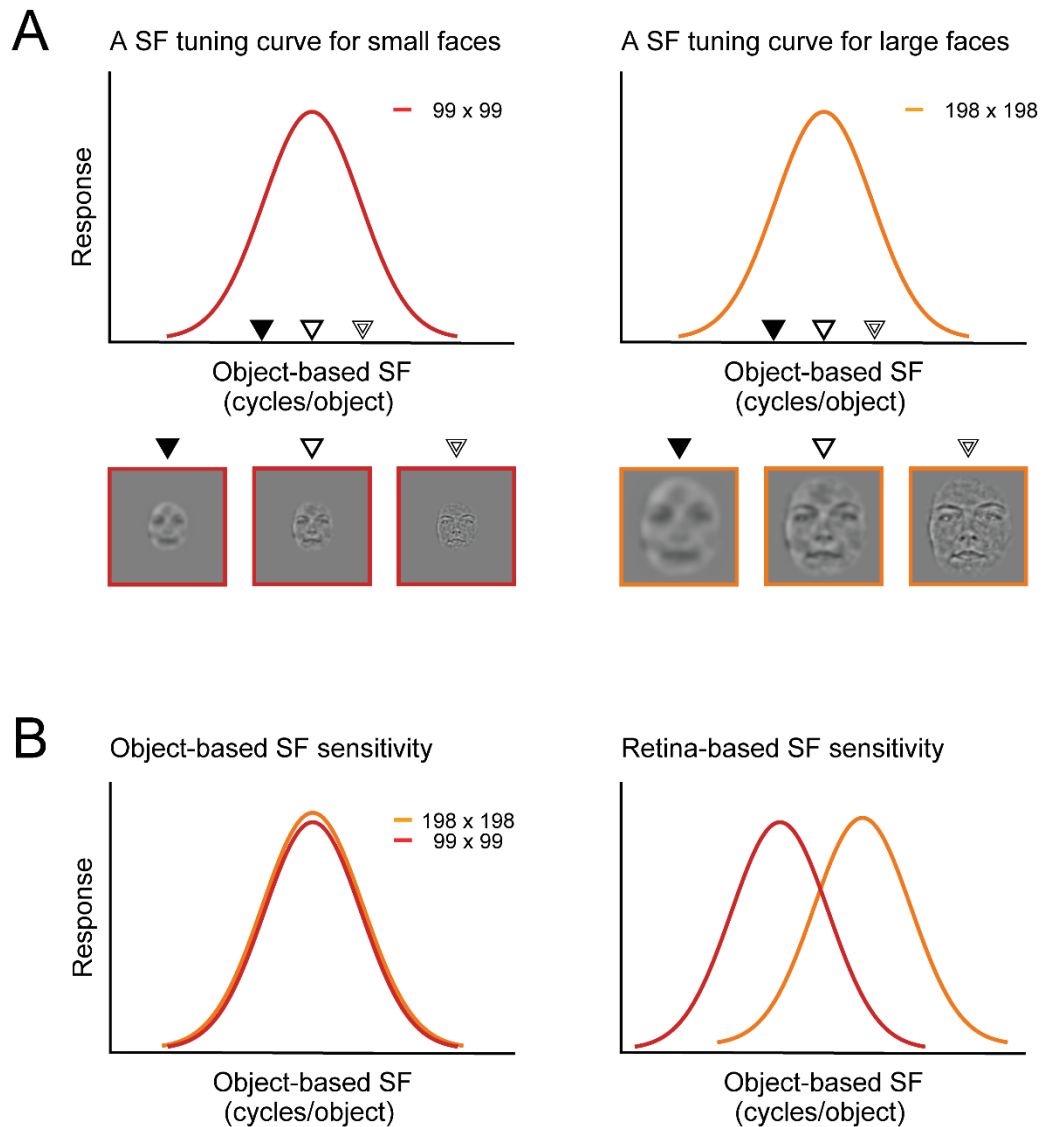


Figure 1.1.2. Object-based SF vs. retina-based SF sensitivity. A. Examples of SF tuning curves. A left curve (red) is an example for small band-pass filtered faces (99×99 pixels). A right one (orange) is that for large faces (198×198 pixels). Same triangles marks images composed of the same center object-based SFs. Object-based SF remains constant regardless of changes in object size. B. Hypothetical tuning curves for object-based SFs of units ideally tuned to object-based SFs (left) or retina-based SFs (right). The tuning properties are characterized by peak shifts (see Materials and Methods). Present examples have ideal peak shifts (left, peak shift = 0; right, peak shift = 1).

Materials and Methods

2.1 Architecture of SNNs

The design and performance of CNN models are primarily influenced by the number of stages and the processing characteristics implemented within these stages. Convolutional filters play an essential role in convolutional processing, whereas window sizes are critical for pooling. With this in mind, we developed SNNs that incorporate the three essential subcortical properties (Fig. 2.1.1B–D; Table 2.1.1). These properties include shallow processing stages, DoG-type organization of receptive fields at the initial processing stage, and larger pooling size.

Unlike typical DNNs, the SNNs consisted of only two sets of convolution and pooling layers followed by two fully connected layers (FC1, FC2), approximating the small number of processing stages of the subcortical pathway. The first convolution layer incorporated 32 DoG-type filters (Fig. 2.1.1C, top) with a spatial resolution of 11×11 pixels, whereas weights in the second convolution layer were initially random, i.e., the filters had no structure, and were gradually changed through training. A rectified linear unit (ReLU) was used as the activation function of a unit in the convolution layers and the first fully connected layer; the ReLU forwards the results of convolution directly to the next stage if they are positive, otherwise it outputs zero. A max pooling operation was performed over 5×5 sliding regions with a stride of 4 for the outputs of convolution layers (Fig. 2.1.1D, top). Max pooling selects the largest value among the responses of units within a sliding window over the preceding convolution layer and forwards the value to the next layer. A local response normalization process was added after the pooling layers to aid generalization (Krizhevsky et al., 2012; we used slightly different parameters from theirs; $k = 1$, $n = 5$, $\alpha = 2 \times 10^{-5}$, $\beta = 0.75$). Every unit in FC1 and FC2 received inputs from all units in

the immediately preceding layer, i.e., each was fully connected. FC1 is the final processing layer, and the seven units in FC2 output the results of processing by the SNNs. These features were implemented to capture the architectural and computational properties of the subcortical pathway, i.e., fewer processing stages compared to the ventral cortical pathway (Fig. 2.1.1A), DoG-type receptive fields in the superficial layer of the superior colliculus (Churan et al., 2012), and large receptive fields of deeper superior colliculus neurons (Wallace et al., 1997). The first three processing layers were intended to represent the superior colliculus, the pulvinar, and the amygdala. The processing types of these layers, i.e., the convolution/pooling in the first two layers and the full connection in FC1 were chosen to match with the retinotopic organization of the three brain regions. The convolution and pooling processes in the first two layers exploit the retinotopy, as the superior colliculus and the pulvinar contain retinotopic maps (Bender, 1981; Chen et al., 2019). The FC1 layer loses the retinotopic information because of the fully connected network, as the amygdala does not have a retinotopic map (Morawetz et al., 2010).

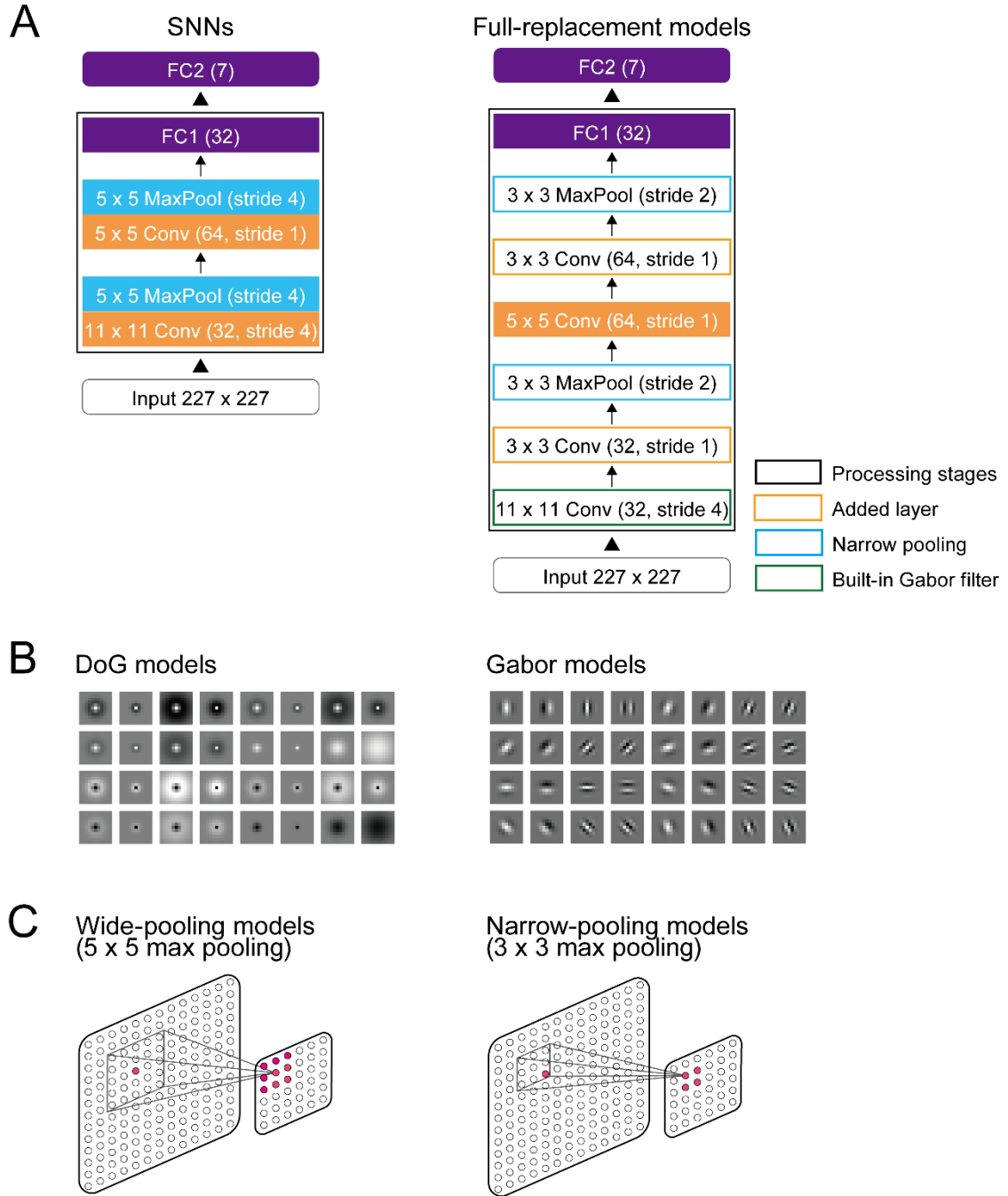


Figure 2.1.1. Shallow neural networks (SNNs) and modifications. **A.** A schematic illustration of the SNNs and full-replacement models. In the full-replacement models, processing layers were added, the filters in the initial layer were changed with built-in Gabor filters, and the range of pooling was narrowed. **B.** DoG filters for the SNNs and DoG models (left), and Gabor filters for the Gabor models (right). **C.** The pooling range for the SNNs and the wide-pooling models (5×5 ; left) and the narrow-pooling models (3×3 ; right).

The SNNs were trained to discriminate images of seven facial expressions: angry, disgusted, fearful, happy, sad, surprised, and neutral (Fig. 2.2.1; see below for details). For each input image, seven units in FC2 yielded scores ranging from 0 to 1 for the seven expression categories, representing the probabilities of classified expressions. The expression with the highest score was taken as the output of the model.

Table 2.1.1 Architecture of the Shallow Neural Network (SNN). Each row describes a layer i with calculation operator F_i , output resolution $H_i \times W_i$, and the number of output channels C_i . Conv denotes convolution layer, and Pool denotes max pooling layer.

Input & Layer i	Operator F_i	Resolution $H_i \times W_i$	#Channels C_i
Input		227×227	3
Layer 1	11×11 Conv (stride 4) & 5×5 Pool (stride 4)	55×55	32
Layer 2	5×5 Conv (stride 1) & 5×5 Pool (stride 4)	4×4	64
Layer 3	Fully connected	1×1	32
Layer 4	Fully connected	1×1	7

The DoG-type filters of the first convolution layer was built by the following formula:

$$DoG(r) = \pm A_1 \exp\left(-\frac{r^2}{2\sigma_1^2}\right) \mp A_2 \exp\left(-\frac{r^2}{2\sigma_2^2}\right), \quad (1)$$

where r is the polar radius from the filter center, A_1 and A_2 are the amplitudes of exponentials of two Gaussian functions, and σ_1 and σ_2 are the standard deviations. Values of A_1 , A_2 , σ_1 , and σ_2 were chosen empirically so that DoG curves took the shapes of Mexican hats. A_1 values were 0.4,

0.67, 0.8, and 1.0. A_2 values were determined based on $A_1 - A_2 = 0.4$. When A_1 was 0.4 (i.e., A_2 is 0, and σ_2 cannot be defined), we set σ_1 values at $1/2\sqrt{2}$, $1/4\sqrt{2}$, $1/8\sqrt{2}$, or $1/16\sqrt{2}$. Otherwise, σ_1 values were $1/2\sqrt{2}$ or $1/4\sqrt{2}$. σ_2 values were based on $\sigma_1/\sigma_2 = 0.2$ or 0.4 (Table 2.1.2). The same number of filters was generated for each A_1 value.

Table 2.1.2 Parameter pairs for DoG-type filters

A_1	A_2	σ_1	σ_2
0.4	0	$1/2\sqrt{2}$	—
0.4	0	$1/4\sqrt{2}$	—
0.4	0	$1/8\sqrt{2}$	—
0.4	0	$1/16\sqrt{2}$	—
$(2/3 + 1) \times 0.4$	$2/3 \times 0.4$	$1/2\sqrt{2}$	$1/\sqrt{2}$
$(2/3 + 1) \times 0.4$	$2/3 \times 0.4$	$1/2\sqrt{2}$	$\sqrt{2}$
$(2/3 + 1) \times 0.4$	$2/3 \times 0.4$	$1/4\sqrt{2}$	$1/2\sqrt{2}$
$(2/3 + 1) \times 0.4$	$2/3 \times 0.4$	$1/4\sqrt{2}$	$1/\sqrt{2}$
$(1 + 1) \times 0.4$	1×0.4	$1/2\sqrt{2}$	$1/\sqrt{2}$
$(1 + 1) \times 0.4$	1×0.4	$1/2\sqrt{2}$	$\sqrt{2}$
$(1 + 1) \times 0.4$	1×0.4	$1/4\sqrt{2}$	$1/2\sqrt{2}$
$(1 + 1) \times 0.4$	1×0.4	$1/4\sqrt{2}$	$1/\sqrt{2}$
$(3/2 + 1) \times 0.4$	$3/2 \times 0.4$	$1/2\sqrt{2}$	$1/\sqrt{2}$
$(3/2 + 1) \times 0.4$	$3/2 \times 0.4$	$1/2\sqrt{2}$	$\sqrt{2}$
$(3/2 + 1) \times 0.4$	$3/2 \times 0.4$	$1/4\sqrt{2}$	$1/2\sqrt{2}$
$(3/2 + 1) \times 0.4$	$3/2 \times 0.4$	$1/4\sqrt{2}$	$1/\sqrt{2}$

I also constructed modified models in which the three properties of the SNNs were replaced one-by-one, two at a time, or all three simultaneously with the corresponding properties in the ventral cortical pathway. First, I added additional convolution layers after the first two to increase the number of processing stages (add-layer model). In adding the convolution layers, the stride of sliding filters was reduced to 1 to keep the output resolutions unchanged before and after adding

the new layers. Also, the new layers contained the same number of filters as the preceding layers to keep the output channels unchanged.

Second, I replaced the DoG-type filters with Gabor-type filters (Gabor model). Gabor-type filtering occurs in simple cells of V1, and emerges in the first layer of DNNs after they are trained to classify object images (Krizhevsky et al., 2012; Rai and Rivas, 2022). I constructed the Gabor-type filters with the following formula:

$$g(x, y; f, \theta) = A \exp\left(-\frac{x^2 + y^2}{2\sigma}\right) \exp(2\pi i f(x \cos \theta + y \sin \theta)), \quad (2)$$

where A is the amplitude of the Gabor filter (i.e., a value of g at $x = y = 0$), i is $\sqrt{-1}$, f is the carrier frequency of a Gabor filter, and θ is the orientation (Movellan, 2008). A was fixed at 0.4 to match the amplitude of the DoG filters. σ was fixed at 1.375 so that the half-amplitude width was half of the filter width. f values were 2 or 4 cycles/image. Orientation θ was 0, 22.5, 45, 67.5, 90, 111.5, 135, or 167.5. I built even- and odd-symmetric filters for every combination of variables. In total, I obtained 32 Gabor-type filters (Fig. 2.1.1B, right).

Finally, I made the convergence field for pooling in the max pooling layers smaller (3×3) than that of the SNNs (5×5), enabling better spatial resolution of processing (Fig. 2.1.1C, right; narrow-pooling model) to mimic the smaller receptive fields in the visual cortices compared to the superior colliculus and pulvinar (Wallace et al., 1997; Van den Bergh et al., 2010; Freeman and Simoncelli, 2011). The pooling range of 3×3 is popular in DNNs (e.g., AlexNet of Krizhevsky et al., 2012; ResNet of He et al., 2016).

2.2 Face images and training of the SNNs

Face images were obtained from Karolinska Directed Emotional Faces (KDEF; Lundqvist et al., 1998) and Radboud Faces Database (RaFD; Langer et al., 2010). Images of the seven expressions of 40 individuals (half females, half males) were chosen from each database (the total number of images was $560 = 7 \times 40 \times 2$). I converted the images from color into grayscale, and extracted the face region by removing hair, neck, and ears with the face-detection function of a computer vision library, OpenCV (Open Source Computer Vision Library; Bradski and Kaehler, 2000). The isolated faces were pasted on a gray background (198×198 pixels; RGB values = 128; Fig. 2.2.1). I augmented the number of face images by changing size and position, and flipping horizontally; seven sizes (28×28 , 56×56 , 85×85 , 113×113 , 141×141 , 170×170 , and 198×198 pixels), five positions (center, left-top, right-top, left-bottom, and right-bottom; directional displacements = 10 pixels), and two horizontally flipped images. The augmentation increased the number of images by 70 times to 39,200. At each training session, I randomly split this augmented set of face images into a training set (29,400 images), a validation set (2,450 images), and a test set (7,350 images). The number of images per facial expression was identical across the stimulus sets. To avoid the inadvertently biased assignment of faces of a particular size, position, or horizontal flip image into a given set, all images from the same individual were assigned into the same set.

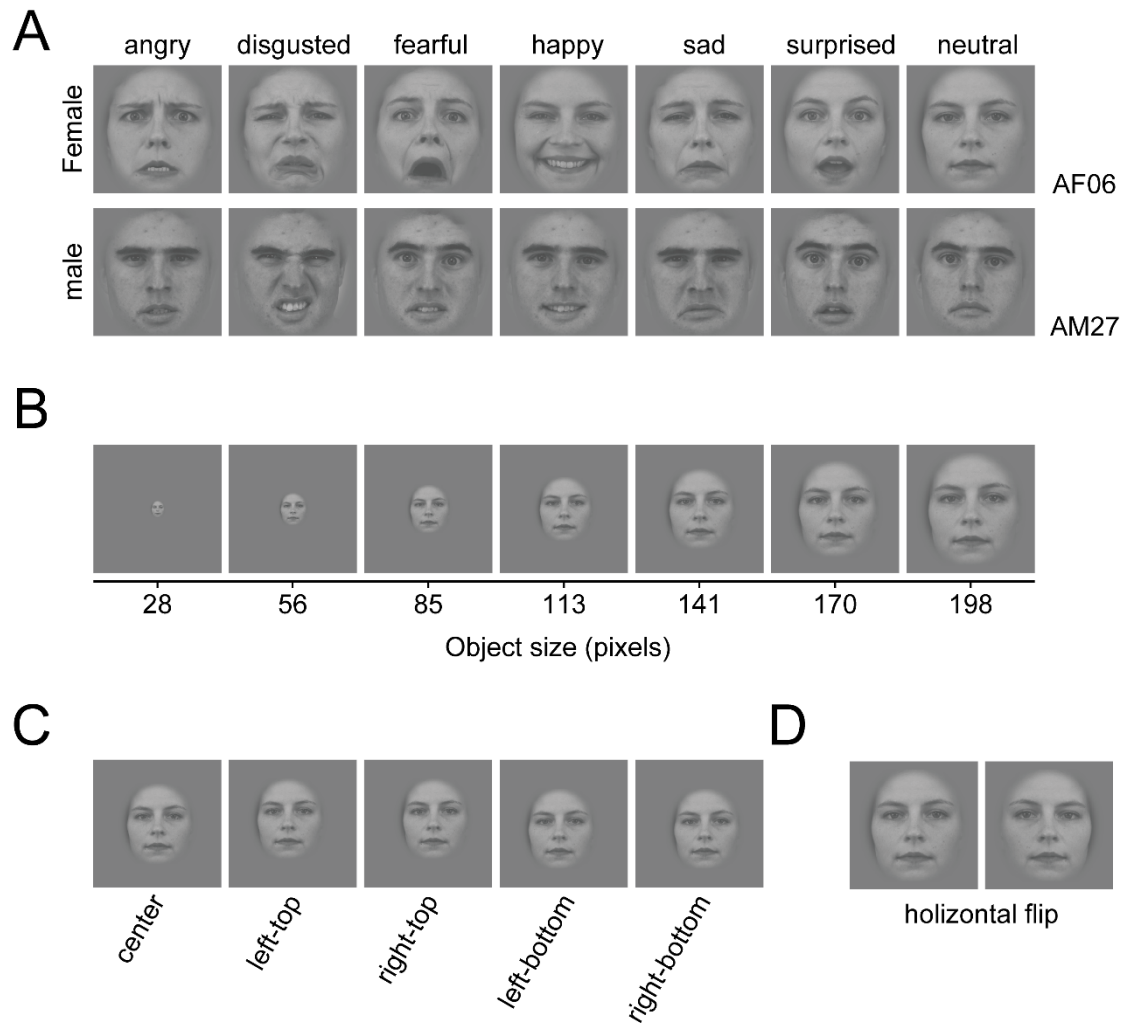


Figure 2.2.1. Face images. A. Examples of presented face images with seven expressions (angry, disgusted, fearful, happy, sad, surprised, neutral) of two individuals (upper: female, AF06; lower: male, AM27). The original images were obtained from the Karolinska Directed Emotional Faces database (Lundqvist et al., 1998). B. Examples of data augmentation with object sizes (28×28 , 56×56 , 85×85 , 113×113 , 141×141 , 170×170 , and 198×198 pixels). C. Examples of data augmentation with position (five positions: center, left-top, right-top, left-bottom, and right-bottom; directional displacements = 10 pixels). D. Examples of data augmentation with horizontal flip.

The training was performed through supervised learning, and was conducted individually 20 times with randomized initial weights except for the built-in weights of the first convolution layer, i.e., we built 20 SNNs that had different initial states. In training, the weights other than the first convolution layer were optimized for classification of face images into the seven categories.

Stochastic gradient descent was used for weight optimization. For each iteration, 32 samples were randomly selected from the training set as a mini-batch. The averaged cross-entropy (Goodfellow et al., 2016) across the 32 images in the mini-batch was calculated as an estimate of loss value, which is a measure of the difference between an estimated value and a true value, and is used for quantifying the training effect. The number of iterations (i.e., weight-updating processes with single mini-batches) was set at 240,000. Initial weight parameters followed a normal distribution with a mean of 0 and a standard deviation of $\sqrt{2/N}$ (N is the total number of weights; He et al., 2015). Weights were updated at each iteration with a constant learning rate of 0.001. This learning rate was determined empirically; a preliminary analysis revealed no decrease in loss values (i.e., no learning) with a learning rate of 0.01, which has frequently been used in the literature (e.g., Simonyan and Zisserman, 2014; Inagaki et al., 2022b). No learnings were also observed with a learning rate of 0.05. A dropout process was added before FC2 to facilitate learning across all units. The proportion of units dropped out of each weight update was set to 0.5. The training was conducted in a Python environment (Chainer 3.0.0; Tokui et al., 2015) on a graphics processing unit (GPU) machine (Intel® Core™ i7-5820K Processor, Intel, Santa Clara, California, USA; The GeForce® GTX 1080 Ti, NVIDIA, Santa Clara, California, USA) with the single-precision floating-point format (float 32, 4 bytes). During our training process, we utilized 417 MiB of GPU memory, with the SNN accounting for 69 MiB of that total. While the SNNs were being trained with the training set, the correct rate and loss value for the validation set were periodically checked

to monitor signs of overfitting. After training was completed, the performance of the models was evaluated using the test dataset. This was done to ensure that the models acquired a genuine ability to classify the facial expressions, as opposed to simply sorting the training images into the seven facial expression categories according to the instruction signals.

2.3 Test for reference frames of SF tuning of model units

A prominent difference in visual responses between the two pathways is the reference frame of neuronal tuning to SFs (Inagaki and Fujita, 2011). Neurons in the inferior temporal cortex, the final stage of the ventral cortical pathway, are tuned to object-based SFs (cycles/object) and represent face patterns in a size-invariant, hence distance-invariant, manner (Fig. 1.1.2B, left). Thus, the ventral cortical pathway converts the representation of SFs in the retina-based coordinate (cycles/degree) to that of object-based SFs. By contrast, many amygdala neurons preserve sensitivity to retina-based SFs. When the stimulus size is changed, they change their preferred object-based SFs; for large stimuli, they respond to higher retina-based SFs, which corresponds to the same retina-based SFs (Fig. 1.1.2B, right). We analyzed the reference frame of FC1 SF tuning to evaluate how well our models captured this characteristic of subcortical processing.

Bandpass-filtered face images were used to examine the SF tunings (Fig. 2.3.1). These images were created by multiplying Gaussian functions with the original face images on the polar Fourier domain. Gaussian functions had 61 different center frequencies between 1 cycle/object and 64 cycles/object. The center frequencies had discrete values at steps of 0.1 cycles/object on a log scale. Gaussian functions shared the same variance at 2.4 octaves, regardless of their center frequencies. The filtered images had amplitude spectra that were determined solely by the

Gaussian function because their spectra were set to be flat before multiplication. To balance the total luminance contrast among the filtered face images, the peak amplitude of the Gaussian function was set inversely proportional to the center image-based SF (Inagaki and Fujita, 2011). These bandpass-filtered images were created for the seven facial expressions at two different sizes (99×99 and 198×198 pixels).

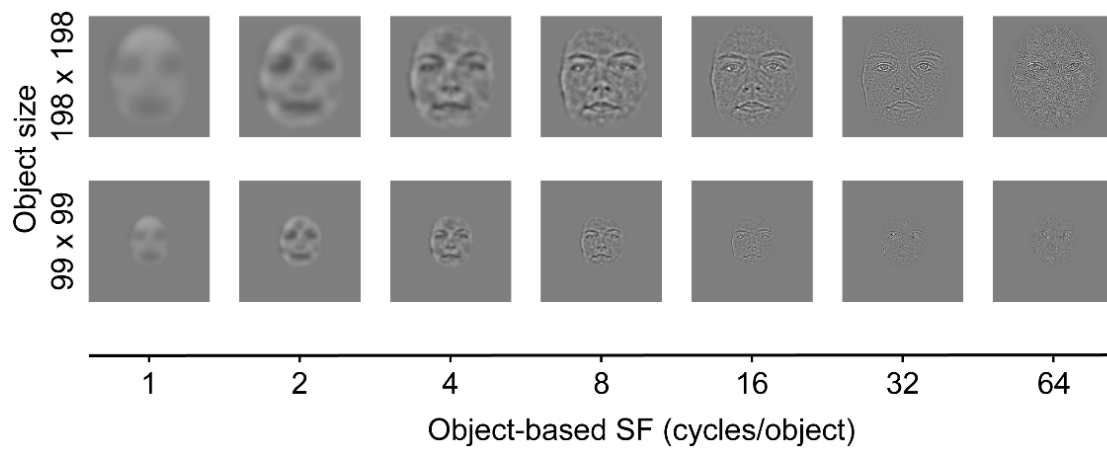


Figure 2.3.1. Test determining whether units are tuned for object- or retina-based spatial frequencies (SFs). The models were fed face images with two different sizes (198×198 and 99×99 pixels) and 64 different bandpass filterings. These images were created by applying two-dimensional bandpass filters that shared the same center object-based SF across different sizes. For each unit in FC1, we obtained responses to images of different center SFs to create tuning curves for object-based SFs. Bandpass filtered faces (for ideal examples of object-based and retina-based SF sensitivities, see Fig. 1.1.2).

The responses of FC1 units to the bandpass-filtered face images were analyzed. To characterize the reference frame of SF tuning of each unit, the peak SFs for the two stimulus sizes were estimated, defined by the SFs at which filtered face stimuli activated a unit most strongly. For a

given unit, 14 peak SFs were determined (two sizes $\times$ seven expressions). The degree to which unit responses to SFs depended on the stimulus size was quantified by calculating differences in peak SFs on a log scale between the two face sizes. A peak shift of 0 indicated that a unit was perfectly tuned to object-based SFs (Fig. 1.1.2B, left). A peak shift of 1 indicated that a unit was perfectly tuned to retina-based SFs (Fig. 1.1.2B, right). This analysis excluded cases in which units did not respond to face images or were not selective for SFs, and cases in which peak SFs were at either end of the tested range of SFs and the peak positions could not be determined.

2.4. Analysis of the effects of the max pooling operation on SF selectivity

The max pooling operation collapses positional information of edges, which is critical for encoding the SF information and detected by DoG and Gabor filters. We therefore examined the effects of the max pooling on the SF selectivity of units. We were particularly interested in the role of the max pooling in converting sensitivity to retina-based SFs to sensitivity to object-based SFs. Bandpass-filtered images of two stimulus sizes (198×198 and 99×99) were fed to the models, and it was determined how different stimulus sizes affected the responses of units to SFs in the first convolution layer (before pooling). Changes in response patterns across the units associated with different stimulus sizes were quantified by calculating the dissimilarity index. The dissimilarity index $D(x, y)$ for responses (x) to large stimuli and those (y) to small stimuli was defined by the Euclidean distance between x and y as follows:

$$D(x, y) = \|x - y\| / (NM), \quad (3)$$

where $\|\cdot\|$ is the Euclidean distance, and N is the number of elements of x and y . M is the maximum value among the 6,405 Euclidean distances calculated for 61 center SFs and the seven facial

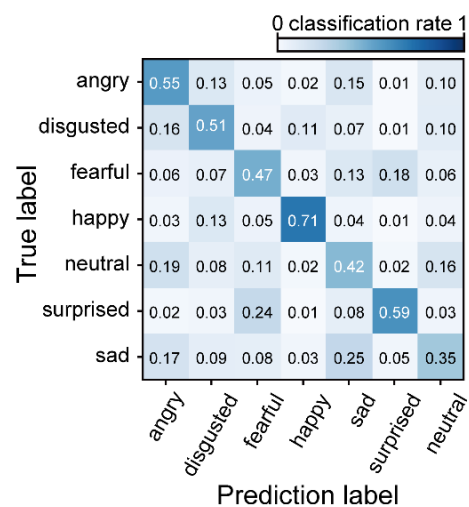
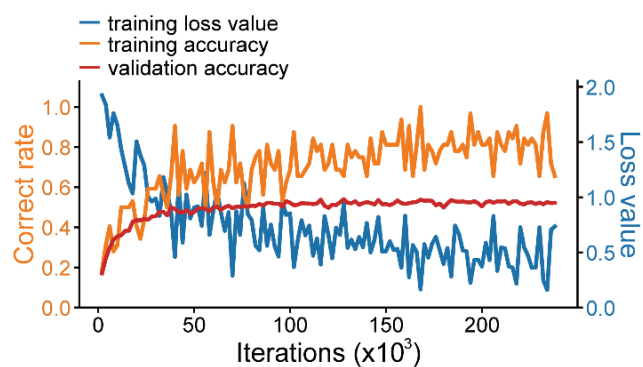
expressions of 15 individuals. To probe the roles of the max pooling, the ratio of the dissimilarity index before pooling in the first convolution layer and after pooling in the first max pooling layer was then calculated. This analysis was applied to the case of wide pooling (5×5) and narrow pooling (3×3), as well as to the case of the SNNs and the Gabor models. By dividing $\|x - y\|$ by M , the dissimilarity index was normalized across the layers (convolution vs. max pooling) and the models (SNNs vs. Gabor models), taking values from 0 to 1.

Results

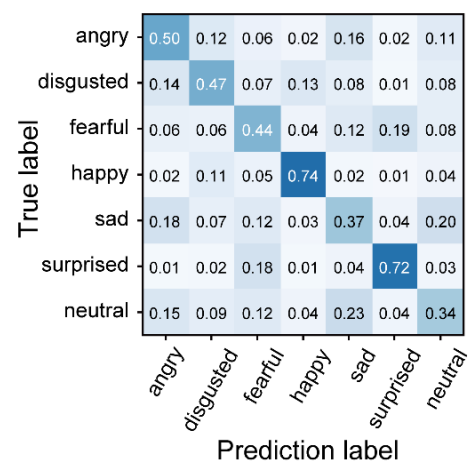
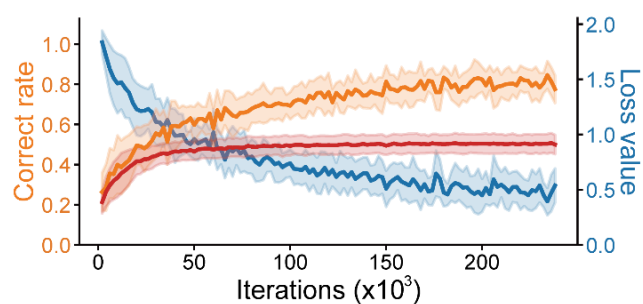
3.1 Performance of SNNs in facial expression classification

The SNNs were trained to classify each face image in the training set into one of the seven facial expressions. The training improved the classification performance rapidly over the initial iterations and then slowly thereafter. The correct rate across the seven facial expressions rose from the chance level (0.14), surpassed 0.6 around 50,000 iterations, further improved to around 0.8 over 150,000 iterations, and reached an asymptote (Fig. 3.1.1A for an example SNN; 3.1.1B for the average of the 20 constructed SNNs; orange lines). The correct rate for the validation set saturated at around 0.5, which was substantially lower than for the training set, indicating insufficient generalization to unseen images. However, the validation correct rate reached a plateau in a similar way to the training correct rate. This indicates that the low correct rate was not the result of inadequate training, but represents the limited learning ability of the SNNs. It also indicates that no overfitting occurred. The loss value also quickly decreased over the initial 50,000 iterations, and became gradually stable (Fig. 3.1.1A, B; cyan lines). The results indicate that within the range of adopted iterations (240,000), the SNNs learned to classify facial expressions without overfitting.

A An example of SNN



B Average of 20 SNNs



C Comparison across facial expressions

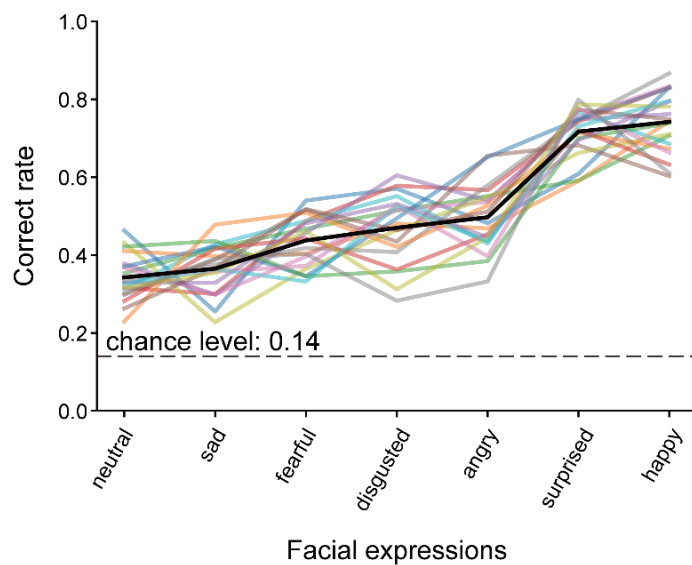


Figure 3.1.1. Learning curves and confusion matrices of an example SNN (A) and the average of 20 SNNs with different initial weights (B). Left panels show changes in correct rates that occurred during training in the training set (orange) and the validation set (red), and loss values (cyan). Confusion matrices on the right indicate the rate of classification of each facial expression (true label) as one of the seven expressions (prediction label). C. The correct rates for the seven expressions across the 20 SNNs. Box plots indicate the median (horizontal line in each box), the 75th and 25th percentiles (upper and lower sides of each box), and the maximum and minimum data except for outliers beyond $\times 1.5$ of the quartile range (whiskers).

The correct rates for the test set were higher than the chance level ($1 / 7 = 0.14$) for all facial expressions, but were modest; the average correct rate across the seven expressions was 0.51 for the example SNN shown in Fig. 3.1.1A. The average correct rate across the 20 constructed SNNs was $0.51 (\pm 0.03, \text{s.d.})$. This was not different from the average correct rate across additional 20 SNNs that were trained with 3,000,000 iterations (0.50 ± 0.03 ; $p = 0.298$, t -test). The training performance thus did not improve even when the SNNs underwent overly excessive training, assuring that the modest correct rate was not due to insufficient training but reflected the limited ability of the SNNs. Confusion matrices showed that the correct rates varied among the facial expressions (Fig. 3.1.1, right panels; $p < 0.01$, ANOVA). Based on the averaged performance, the classification performance of the SNNs was best for happy (0.74) and surprised faces (0.72), followed by angry (0.50), disgusted (0.47), and fearful (0.44) faces, and was worst for sad (0.34) and neutral (0.34) faces. Sad faces were often confused with neutral, fearful, and angry faces. Neutral faces were often confused with sad and angry faces. This pattern was largely consistent across the 20 constructed SNNs (Fig. 3.1.1C; $p < 0.001$ for expressions, $p = 0.562$ for models, two-way ANOVA).

3.2 Effects of modification of SNNs on classification performance

Replacement of any one or a combination of the three subcortical properties, the shallowness of processing stages, DoG-type receptive fields at the initial stage, and spatial pooling over a wider visual field, with the corresponding cortical properties improved the classification accuracy (Fig. 3.2.1A; $p < 0.01$, ANOVA). The correct rates averaged across the seven expressions and the 20 constructed models of each modification were increased from 0.51 in the SNNs to 0.55 in the narrow-pooling models, 0.64 in the Gabor models, and 0.69 in the add-layer models ($p < 0.01 / 28=8C_2$; t -test with Bonferroni correction). Among these models with one replaced property, the add-layer models exhibited the best performance. The results indicate that all three properties had effects on classification performance, and the layer structure was the most influential.

When two properties were replaced together, the narrow-pooling + Gabor models and the narrow-pooling + add-layer models performed better than the narrow-pooling models and the Gabor models ($p < 0.01 / 28=8C_2$; t -test with Bonferroni correction), but comparably to the add-layer models (correct rates, 0.69, $p = 0.778$ for the narrow-pooling + Gabor models; 0.69, $p = 0.564$ for the narrow-pooling + add-layer models). The Gabor + Add-layer models performed better than all one-property-replacement models (correct rate, 0.74; $p < 0.01 / 28=8C_2$). When all three properties were replaced together (full-replacement models), the correct rate was 0.77, which was better than all other models ($p < 0.01 / 28=8C_2$) except for the Gabor + add-layer models ($p = 0.00846$). The performance was improved for all facial expressions (Fig. 3.2.1B; mean correct rates: happy = 0.92; surprised = 0.85; disgusted = 0.81; neutral = 0.78; angry = 0.75; fearful = 0.66; sad = 0.65). As in the SNNs, it was highest for happy and surprised faces, and lowest for fearful and sad faces. The performance was improved most for neutral faces (SNNs, 0.37; full-

replacement models, 0.78). The variance of the correct rates was affected both by facial expressions and models (Fig. 3.2.1C; $p < 0.001$ for facial expressions, $p = 0.00285$ for models, two-way ANOVA). The improved performance of the two-property-replacement and full-replacement models indicate that the effects of the three features on classification performance were partially additive, suggesting that the three features exerted their effects partially independently.

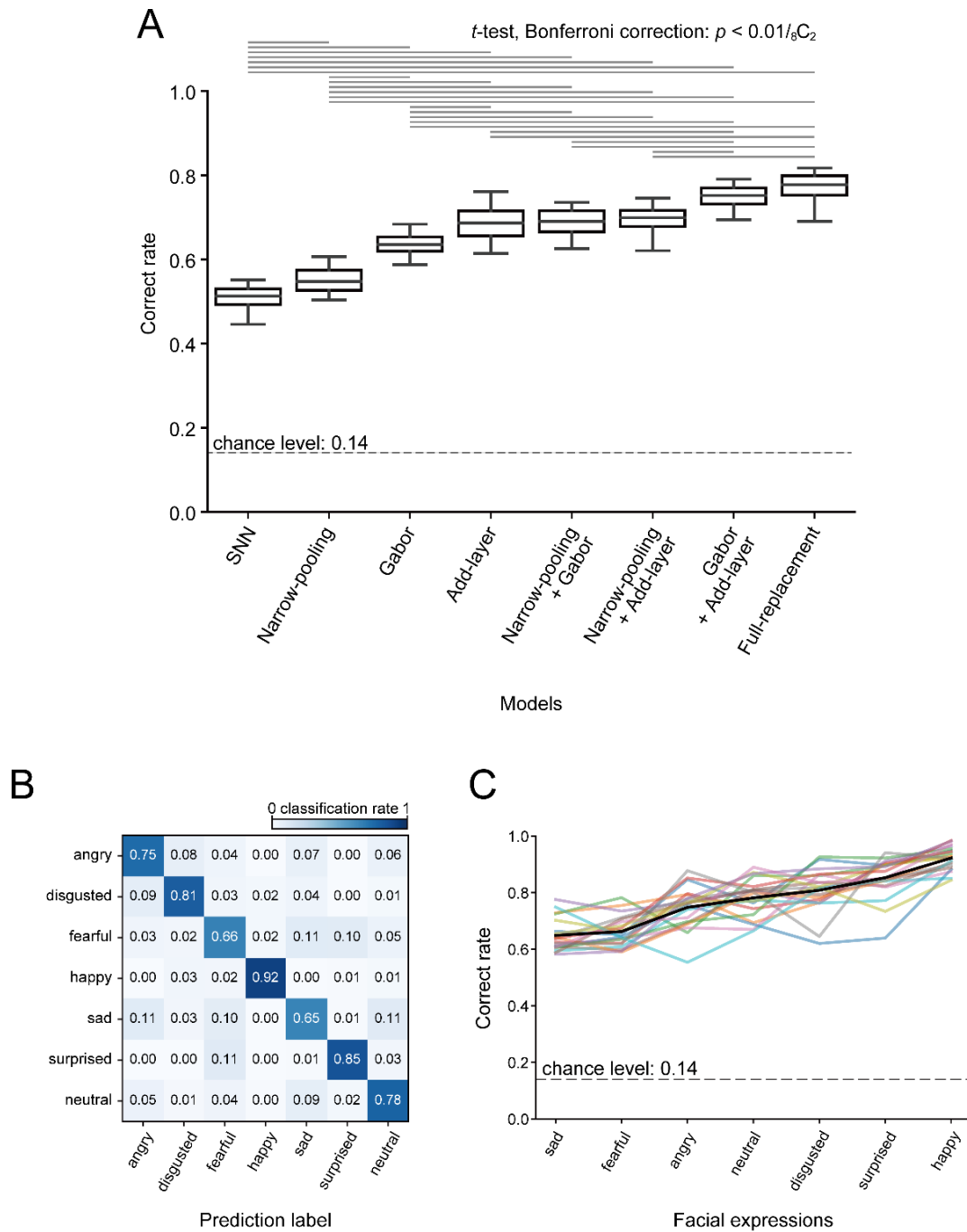


Figure 3.3.1. Effects on model performance of replacing subcortical properties with corresponding cortical properties. (A) Discrimination performances of the SNNs and the modified models. The discrimination performance differed across the models (ANOVA, $p < 0.001$). The pairs of models with statistically significant differences in performance are linked with horizontal

lines in the upper part (t-test, Bonferroni correction, $p < 0.01/28=8C2$). (B) Confusion matrix for the full-replacement models (average of the 20 constructed models). (C) The correct rate for the seven expressions across the 20 full-replacement models. The black line indicates the mean, and lines with other colors indicate the data for individual full-replacement models.

3.3 Spatial frequency representation in the FC1 layer

As shown above, the SNNs exhibited modest performance in facial expression classification, and this performance was improved by changing SNN subcortical properties to corresponding cortical properties. These findings suggest that the SNNs captured aspects of processing in the subcortical pathway to the extent that they explained the suboptimal perceptual performance of V1-lesioned patients. We next looked into individual computational units to gain insights about the processing in the models. I examined SF sensitivities of FC1 units by using two different sizes of input images (198×198 and 99×99). This procedure allowed us to determine whether units were sensitive to object-based or retina-based SFs (Fig. 1.1.2; see Materials and Methods).

FC1 units of the SNNs exhibited a variety of dependencies of SF tunings on stimulus size (Fig. 3.3.1A). Some units responded to the same range of SFs for both large and small stimuli, and the peak positions of the SF tuning curves remained unchanged (Fig. 3.3.1Aa, Ab). Other units exhibited different preferred SFs for large and small stimuli, and in these cases the peak position shifted horizontally along the abscissa (Fig. 3.3.1Ac–f). We quantified these shifts by measuring the difference between preferred SFs expressed in logarithm for the two stimulus sizes. A peak shift of 0 means that the unit responded to object-based SFs, whereas a peak shift of 1 means that

the unit responded to retina-based SFs. The peak shifts of the example units shown in Fig. Fig.

3.3.1A were 0.0 (a), 0.1 (b), 0.1 (c), 0.9 (d), 1.0 (e), and 1.3 (f).

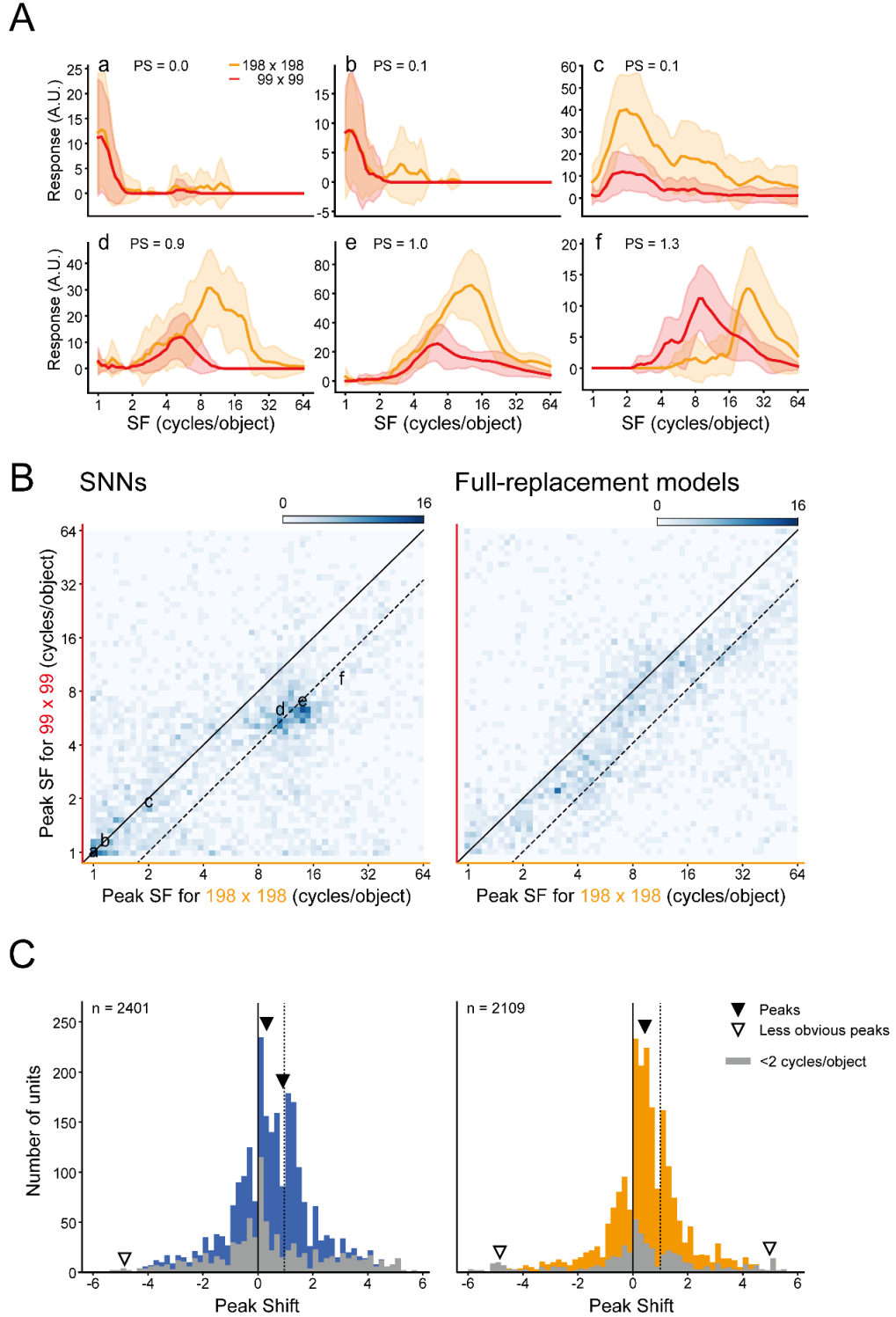


Figure 3.3.1. SF tuning reference frames of FC1 units of the SNNs and full-replacement models. Responses of FC1 units to SF-filtered face images were examined at two different sizes (198×198 , 99×99 pixels). (A) Six example FC1 units of SNNs with a different peak shift (PS). (B) Two-dimensional histograms of peak SFs at large images versus small images for the SNNs (left) and the full-replacement models (right). Solid lines indicate peak shifts of 0, and dashed lines indicate peak shifts of 1. (C) Distribution of peak shifts of units in the 20 SNNs (left) and the 20 full-replacement models (right). Arrowheads indicate the estimated locations of multiple peaks in the distribution (solid: major peaks, open: statistically detected but less obvious peaks). Gray columns indicate units with a response peak at SFs below 2 cycles/object for large and/or small stimuli.

I plotted the peak positions of 2,401 FC1 units of the 20 SNNs in a two-dimensional space defined by the peak SF for the large stimuli on the abscissa, and the peak SF for the small stimuli on the ordinate (Fig. 3.3.1B, left). Note that 46% of FC units were excluded from this analysis, either because they were not sensitive to SFs (23%) or because the largest responses were found at the end of the examined range of SFs and the peak SFs could not be determined (23%). The diagonal solid line in Fig. 3.3.1B represents the responses with a peak shift of 0, and the dashed line next to it represents the responses with a peak shift of 1. FC1 units of the SNNs were distributed in multiple groups in this scatter plot. One prominent group was selective to low SFs and was centered on the diagonal, i.e., peak shift values around 0. Another group was selective to higher SFs, and was diffusely distributed along the dashed line indicative of peak shift values around 1. The multimodality of the distribution can be seen more clearly in the histogram (Fig. 3.3.1C, left). We applied an excess mass test for multimodality (Ameijeiras-Alonso et al., 2019, 2021) to this distribution. The null hypothesis of the statistical test is that the true number of peaks is N ($N = 1$,

2, 3, ...). Repeating the test with an increment of N , the number of peaks of the distribution is estimated as the smallest N under which the null hypothesis is not rejected. The alternative hypothesis is that the true number of peaks is greater than N , because null hypotheses with the number of peaks $< N$ have already been rejected. The test revealed that there were three peaks in the distribution of the SNNs (first p -value < 0.001 , second p -value < 0.001 , third p -value $= 0.114$). Based on the probability density function derived from the histogram (Ameijeiras-Alonso et al., 2021), the peaks were estimated to be located at -4.85 , 0.144 , and 0.909 (open and solid triangles in Fig. 3.3.1C, left). Comparing this result and the density map, the peak at 0.144 was from the low SF group and the peak at 0.909 was from the high SF group. Although the third peak at the far periphery (-4.85 , open triangle) was statistically detected, it was negligibly small in height compared to the other two peaks (1.1% of the peaks near 0 and 1). The results indicate that FC1 contained two major groups of units, those sensitive to retina-based SFs (peak shifts around 0) and those sensitive to object-based SFs (peak shifts around 1).

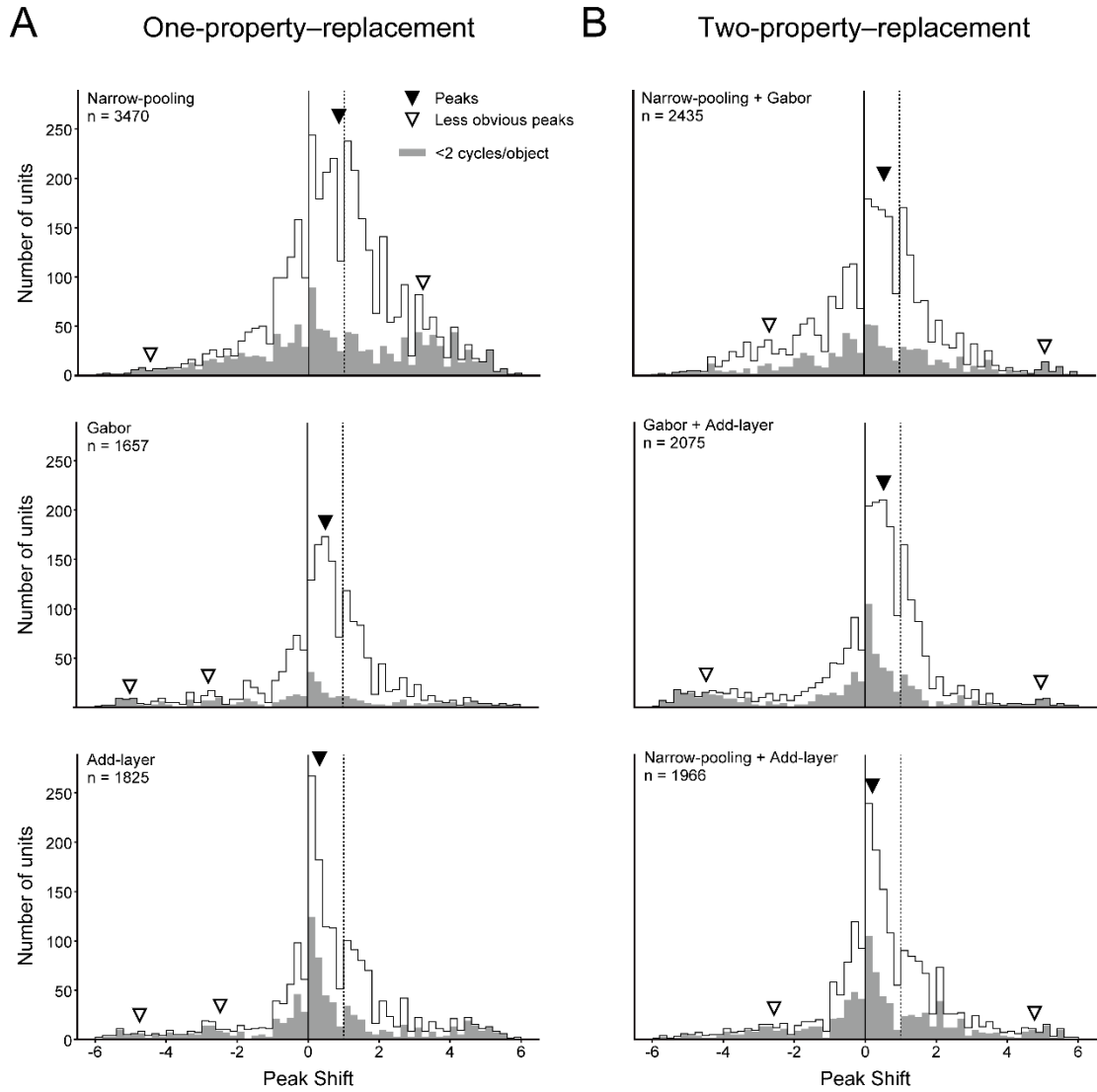


Figure 3.3.2. Distributions of peak shifts of FC1 units of the SNNs and one-property or two-property replacement models. (A) Data from the narrow-pooling model, add-layer model, and Gabor model. (B) Data from models with two modifications: narrow-pooling + Gabor, Gabor + add-layer, and narrow-pooling + add-layer. Gray columns indicate units with a response peak at SFs below two cycles/object for large or small stimuli. Arrowheads indicate the estimated locations of multiple peaks in the distribution (solid: major peaks, open: statistically detected but small peaks). Gray columns indicate units with a response peak at SFs below 2 cycles/object for large and/or small stimuli.

The distribution of peak shift values was drastically altered in the full-replacement models. In the scatter plot shown in Fig. 3.3.1B (right), most data points were continuously distributed in an elongated area between the diagonal and dashed lines, indicating that the SF reference frame of most units was intermediate between retina-based and object-based. An excess mass test resulted in a detection of three peaks in the distribution at -4.84 , 0.436 , and 4.98 (solid and open triangles; first p-value < 0.001 , second p-value < 0.001 , third p-value $= 0.096$). The second and third peaks at the far periphery at -4.84 and 4.98 (open triangles) were much smaller in height than the primary peak at 0.436 (3.1% and 2.6% of the primary peak, respectively), making the distribution nearly unimodal.

Given the unimodality of the major peak in the full-replacement models, we analyzed one- and two-property replacement models to determine which subcortical properties were essential for the multimodal distribution of the SNNs. All these models exhibited unimodal distributions of the major peak (solid triangles) at different peak positions. An excess mass test for multimodality detected two other less obvious peaks (open triangles) in each model (Fig. 3.3.2A, B; Table 3.3.1) as in the cases for the SNNs and the full-replacement models (Fig. 3.3.1C). These smaller peaks were 2.6–26 % of the major peaks in their height, and found at the far periphery of the distribution (Table 3.3.2).

Table 3.3.1 Estimated peak locations by excess mass tests for multimodality

model	1st peak; p -value	2nd peak; p -value	3rd peak; p -value
Narrow-pooling	0.849; $p < 0.001$	-4.46; $p < 0.001$	3.23; $p = 0.094$
Gabor	0.515; $p < 0.001$	-5.00; $p < 0.001$	-2.79; $p = 0.064$
Add-layer	0.324; $p < 0.001$	-4.75; $p < 0.001$	-2.48; $p = 0.012$
Narrow-pooling + Gabor	0.564; $p < 0.001$	5.09; $p < 0.001$	-2.67; $p = 0.09$
Narrow-pooling + Add-layer	0.195; $p < 0.001$	4.77; $p < 0.001$	-2.57; $p = 0.0096$
Gabor + Add-layer	0.520; $p < 0.001$	-4.48; $p < 0.001$	4.96; $p = 0.028$

Table 3.3.2 Estimated value of the density by excess mass tests for multimodality

model	1st peak	2nd peak	3rd peak
Narrow-pooling	0.299	0.00903	0.0781
Gabor	0.405	0.0163	0.0295
Add-layer	0.341	0.0138	0.0297
Narrow-pooling + Gabor	0.319	0.0136	0.0545
Narrow-pooling + Add-layer	0.400	0.0222	0.0309
Gabor + Add-layer	0.432	0.0349	0.0113

Each of the three one-property replacement models showed characteristic distribution of the peak shift values. The narrow pooling models contained units with peak shift values between 0 and 1 in addition to units with peak shift values around either 0 or 1. The distribution became unimodal

and broad, and was estimated to be centered at 0.85. In the Gabor models, units with peak shift values intermediate between 0 and 1 were the most abundant with a smaller number of units of peak shift values around 0 and 1. The distribution peak was estimated at 0.52. In the add-layer models, units with peak shift values around 0 dominated and exhibited a sharp peak at 0.32 in the distribution. These units were mostly sensitive to low SFs (< 4 cycles/object; gray columns). As to the two-property replacement models, the narrow-pooling + Gabor models and the Gabor + add-layer models showed a broad distribution straddling the peak values from 0 to 1 (peak for the former, 0.56; peak for the latter, 0.52), whereas the narrow-pooling + add-layer models showed a sharp distribution peak at 0.20. The results indicate that all of the three computational properties were responsible for the multimodal distribution of peak shift values observed in the SNNs. Particularly, the reduction of the number of units with peak shift values around 1 in the Gabor models and the add-layer models suggests that the shallowness and the DoG-type filters were necessary for preserving the unit sensitivities to retina-based SFs. The broad distribution observed for the narrow-pooling models and the narrow-pooling + Gabor models suggests that the wide pooling employed in the SNNs contributed to the two peaks at 0 and 1, by preventing creation of units with peak shift values intermediate between 0 and 1.

3.4 Training and testing of SNNs with a different face database

Next, I explored the generalizability of our findings based on the KDEF and RaFD databases by assessing the SNNs trained with another face database, KRC. I trained and tested an additional set of 20 SNNs using the KRC database, following a similar approach to the main analysis with the KDEF and RaFD databases. Similar to the KDEF/RaFD-trained SNNs, the SNNs trained on the KRC database achieved modest correct rates above chance level for all facial expressions. The mean correct rate across the seven facial expressions was 0.49 (s.d. = 0.02), slightly lower

but still comparable to 0.51 in the KDEF/RaFD-trained SNNs. Furthermore, the performance order for the seven expressions was largely consistent with the models trained on the KDEF/RaFD databases: the performance was best for happy (0.76) and surprised (0.71) faces, moderate for disguised (0.44), fearful (0.42), neutral (0.42), and angry (0.41) faces, and poor for sad (0.28) faces (Fig. 3.4.1A).

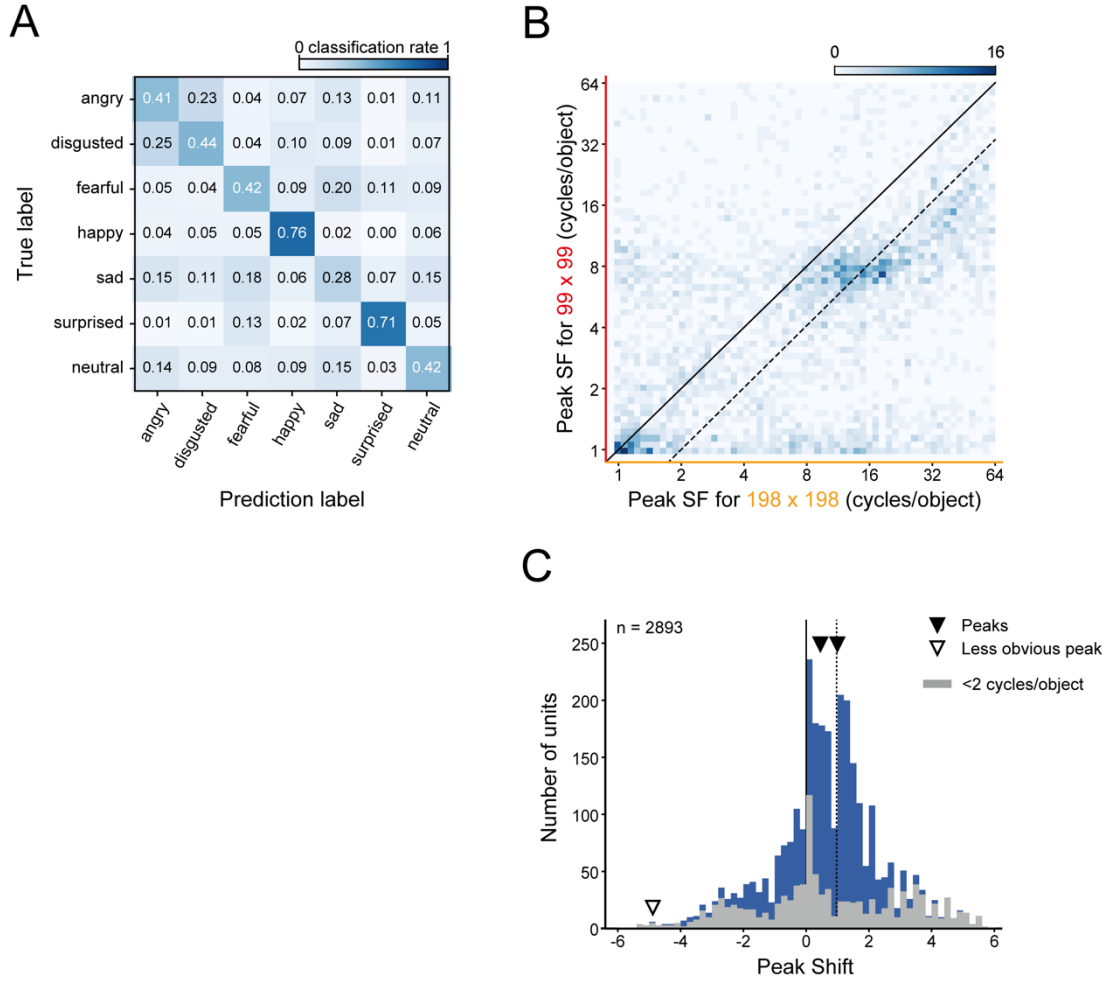


Figure 3.4.1. Effects of alternation of training and test datasets to the KRC, a Japanese facial expressions dataset, on the classification performance of 20 SNNs (A) and the SF tuning reference frame of FC1 units in these networks (B and C). The average performance across the 20 SNNs was obtained in the same way as in Fig. 3.1.1 and responses of FC1 units were obtained in the same way as in Fig. 3.3.1. (A) A confusion matrix illustrates the rate of classification of each facial expression (true label) as one of the seven expressions (prediction label). (B) A two-dimensional histogram of peak SFs obtained with large (198×198) versus small (99×99) stimulus images. (C) Distribution of peak shifts of units. Arrowheads indicate the estimated

locations of multiple peaks in the distribution. Gray columns indicate units with a response peak at SFs below 2 cycles/object for large and/or small stimuli.

Analysis of the SF representation in the FC1 layer revealed that, similar to the KDEF/RaFD-trained SNNs (Fig. 3.3.1B, left), the KRC-trained SNNs exhibited two major groups of units. One group encoded object-based SFs in a low SF range, while the other encoded the retina-based reference frame in a higher SF range (Fig. 3.4.1B). The distribution of units in two-dimensional histograms was similar between the KRC-trained SNNs and the KDEF/RaFD-trained SNNs (Pearson's correlation coefficient, $r = 0.42$, $p < 0.001$). The multimodality test revealed that the distribution of the peak shift had three peaks, located at -4.89 , 0.455 , and 0.998 (the first and second p-values < 0.001 , the third p-value $= 0.12$; excess mass test for multimodality; solid and open arrowheads in Fig. 3.4.1C). The units sensitive to low SFs below 2 cycles/object were most frequently observed at a peak shift of approximately 0 (gray columns), which is also similar to the observations from KDEF/RaFD databases.

The results suggest that my main findings, the modest classification performance and the bimodal distribution of units based on the SF reference frame, are applicable across the employed face databases, even those comprising faces of diverse ethnic backgrounds.

3.5 Effects of max pooling on SF tuning

I showed above that FC1 units of the SNNs were roughly grouped into two populations in terms of the reference frame of SF encoding. Because the max pooling yields the same output from a population of convolution layer units in response to slightly different retina-based SFs, the max

pooling operation is likely to affect the encoding of SFs. Therefore, I examined the effect of max pooling on the representation of SFs.

I first analyzed the responses in across the 96,800 units ($32 \text{ filters} \times 55 \times 55 \text{ resolution}$) in the first convolution layer. We obtained these response patterns by feeding SF-filtered faces of two sizes (198×198 and 99×99 pixels; Fig. 2.3.1) to the models, and quantified the difference between the SF tunings obtained by the two stimulus sizes by calculating the dissimilarity index (see Materials and Methods). In the SNNs with DoG filters, the dissimilarity index was high (around 0.6) for a low SF range up to approximately 4 cycles/object, but gradually decreased over a higher range of SFs (Fig. 3.5.1A, black curve). In the pooling layer, the dissimilarity index of the 6,272 units ($32 \text{ filters} \times 14 \times 14 \text{ resolution}$) became lower for a low SF range of <4 cycles/object than that of the convolution layer. For a high SF range, >4 cycles/object, by contrast, it became higher than that of the convolution layer (Fig. 3.5.1A, compare orange and cyan curves with the black curve). Thus, max pooling resulted in the SF tuning becoming similar between the two stimulus sizes for a low SF range, consistent with the results of peak shift analysis (Fig. 3.3.1C). Although these changes were observed both for wide pooling (5×5) and narrow pooling (3×3), the effects were larger for the former than for the latter (Fig. 3.5.1A, compare the orange curve with the cyan curve). This was evident when we plotted the ratio of dissimilarity indices between before and after pooling (Fig. 3.5.1B). Furthermore, the ratio curve for wide pooling had smaller standard deviations than that for narrow pooling (shown as shades in Fig. 3.5.1B), indicating that wide pooling exerted its effects more consistently across the 15 individual faces and the 7 facial expressions than narrow pooling.

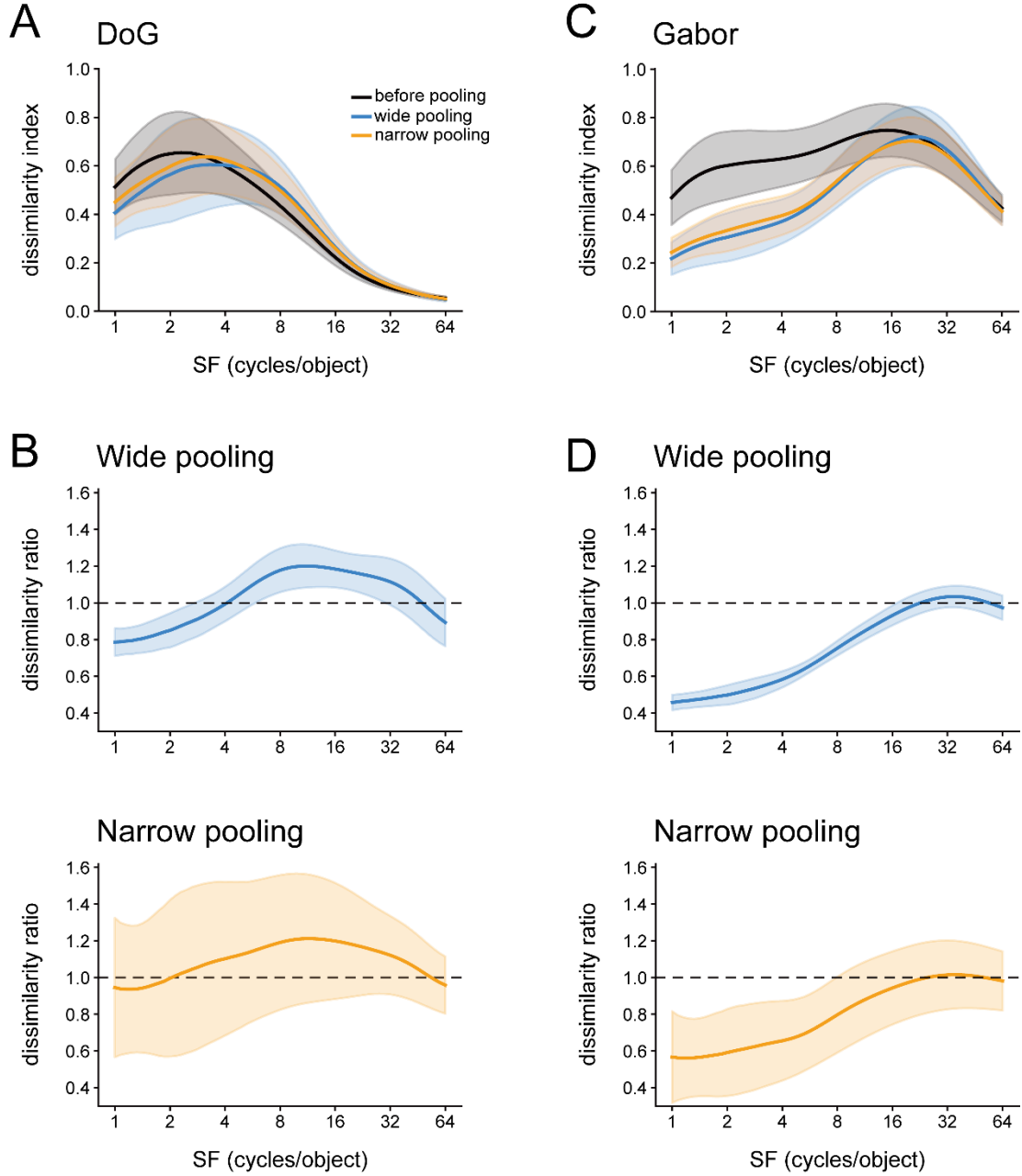


Figure 3.5.1. Effects of max pooling on the size-invariant responses to SFs. (A) Dissimilarity index curves of responses of units in the first convolution layer of the SNNs before max pooling operation (black), after wide pooling (blue), and after narrow pooling (orange). Dissimilarity indices, defined by the Euclidean distances of unit responses between different stimulus sizes (see Materials and Methods), are plotted against the center SFs of input images. Solid lines indicate

the means of dissimilarity indices across the seven facial expressions. Shades indicate standard deviations. Each dissimilarity index was normalized by the number of units and the maximum values. (B) Dissimilarity ratios of inputs and outputs of the max pooling operation (upper, after wide pooling; lower, after narrow pooling). (C, D) Data from units in the first convolution layer of the Gabor models. The conventions are the same as in A and B.

By contrast, the convolution-layer units of the Gabor models exhibited a constantly high dissimilarity index over most of the SF range (Fig. 3.5.1C, black). However, when we applied max pooling with windows of either 5×5 or 3×3 in size, the dissimilarity index became small over the entire SF range, with the largest decrease for 1–16 cycles/object (Fig. 3.5.1C; compare orange and blue curves with the black curve). As in the case of the SNNs, the effect was stronger for wide pooling than for narrow pooling (Fig. 3.5.1D). These results demonstrated that max pooling rendered the SF tuning more invariant to stimulus size for units sensitive to low SFs, enabling them to represent SFs in the object-based coordinate. Regardless of the filter type in the first convolution layer (i.e., DoG vs. Gabor), wide pooling was more effective than narrow pooling in creating this response property.

3.6 Mathematical formulation of the max pooling operation

I next developed a formal description of max pooling to explain its effects of size invariance on the unit responses to low SF images. Low SF images only contains some simple shapes. Simple shapes are characterized by their edges which are simple closed curves. They can be considered as Jordan curves, naturally inducing the area of inner regions surrounded by them. Thus, size invariance to low SF images can be simplified as the area invariance of regions surrounded by edges. In the following section, I provide a proof that the max pooling operation have effects of

size invariance by demonstrating that it does not change outputs against some changes of inputs and there exists curves such that are converted into the same outputs by the max pooling but have different area of inner regions on the input space. While this statement seems obvious from point of view that max pooling ignores slight differences of positions of signals, it takes some technical setup and ingenuity to prove it by elementary means.

The max pooling calculates the maximum values of inputs within regions of interest (ROI) and assigns a vector whose elements are equal to maximum values of individual ROIs. Thus, the max pooling operation can be considered as a mapping from a topological space (defined by outputs of the first convolution layer) to an n -dimensional space (where n is the number of regions of interest) in the max pooling layer. In the proof, I provide the max pooling assigns the same vector to Jourdan curves surrounding different area.

Let X be a topological space. The max value calculation φ is defined by the following:

$$\begin{array}{ccc} \varphi: \mathcal{O}_{X \times \mathbb{R}} & \rightarrow & \mathbb{R} \\ \Downarrow & & \Downarrow \\ A & \mapsto & \varphi(A) = \sup \bar{h}, \end{array} \quad (4)$$

where $\mathcal{O}_{X \times \mathbb{R}}$ is a product topology of $\mathcal{O}_X$ and the $\mathbb{R}$ and $\bar{h}$ is a height function expanded as it be cumulative:

$$\begin{array}{ccc} \bar{h}: \mathcal{O}_{X \times \mathbb{R}} & \rightarrow & 2^{\mathbb{R}} \\ \Downarrow & & \Downarrow \\ A & \mapsto & \varphi(A) \end{array} \quad (5)$$

$$\varphi(A) = \{z | x \in \pi(A), z = \sum_{y \in h(A|_x)} y\}, \quad (6)$$

where π is a natural projection from $X \times \mathbb{R}$ to X , h is height function, and $A|_x$ is $\{a \in A | \pi(a) = x\}$. As a special case, define $\bar{h}(\emptyset)$ as 0. When A contains only one height to each point of $\pi(A)$, $\bar{h}(A) = h(A)$. It is necessary for defining heights of the inputs as curves passing across same points more than twice. As a special case of φ , define $\varphi(\emptyset)$ as 0. Let $\{U_k\}_{k=1}^n$ be a family of compact sets and $X = \bigcap_{k=1}^n U_k$. It is a collection of ROIs and n is the number of that regions. Then, a pair of $(\varphi, \{U_k\})$ fixes a max pooling operation F as

$$\begin{array}{ccc} F: & \mathcal{O}_{X \times \mathbb{R}} & \rightarrow \mathbb{R}^n \\ & \Psi & \Psi \\ & A & \mapsto \begin{bmatrix} \varphi(A|_{U_1}) \\ \vdots \\ \varphi(A|_{U_n}) \end{bmatrix}, \end{array} \quad (7)$$

where $A|_{U_k}$ is $\{a \in A | \pi(a) \in U_k\}$. For any $\{U_k\}$, F gives an equivalence relation between some paths. $p: [0, a] \rightarrow X \times \mathbb{R}$ and $p': [0, a'] \rightarrow X \times \mathbb{R}$ where a and a' are positive real numbers.

This operation provides an equivalence relation with

$$p \sim p' \stackrel{def}{\iff} F(p([0, a])) = F(p'([0, a'])). \quad (8)$$

Defined by the equality of n -dimensional vectors, it clearly satisfies the definition of equivalence.

To define area (i.e., 2D volume), I set X as a compact convex set in $\mathbb{R}^2$ and set the height as $[0, \infty)$.

A compact subset in $\mathbb{R}^2$ is a Lebesgue-measurable (i.e., that area in $\mathbb{R}^2$ is measurable) and it is

natural to assume that inputs are 2D-images and outputs are rectified in standard CNNs. The compact convex set is a generalization of 2D-image because 2D- image have a finite size, no hole, and no dent. Compact equips a set with "finite size" and "no hole" and convex does it with "no dent". Let p, q be loops in $X \times [0, \infty)$. If an area of inner region bounded by $\pi(p)$ differs from that bounded by $\pi(q)$ but $p \sim q$ with the max pooling operation, I can say the max pooling operation has the size invariance. For simplicity, I deal with only loops of which projection to “base” are Jordan curves.

Definition 1 (Lifted Jordan curve) *A lifted Jordan curves is a path in $\mathbb{R}^2 \times [0, \infty)$ of which a projection to $\mathbb{R}^2$ is a Jordan curve.*

Definition 2 (CNN space) *Let X be a compact convex set of $\mathbb{R}^2$. A CNN space is a topology on $X \times [0, \infty)$ induced by $\mathbb{R}^3$*

I show the size invariance of max pooling operation from the following theorem;

Theorem 1 *Let S be a set of lifted Jordan curves in a CNN space and $[p]$ be a equivalence class given by $p \in S$ and the max pooling equivalence relation. Areas of inner region bounded by $\pi(p)$ ($p \in [p]$) are bounded by Areas of inner regions bounded by curves whose projections by π are composed of boundary points of ROIs.*

Proof. Let Λ be an index set of the ROIs such that $\{\lambda \in \Lambda | \pi(p) \cap U_\lambda \neq \emptyset\}$. Then, $\{U_k\}_{k=1}^n = \{U_\lambda\}_{\lambda \in \Lambda} + \{U_\mu\}_{\mu \in \Lambda^c}$ where $\Lambda^c = \{1, 2, \dots, n\} \setminus \Lambda$. Denoting $\bigcap_{\lambda \in \Lambda} U_\lambda$ as U_+ and $\bigcap_{\mu \in \Lambda^c} U_\mu$ as U_0 , define $V = U_+ \setminus U_0$. V means a region which lifted Jordan curves can pass across with preserving

the equivalence relation with p . If a curve pass across the complement of V , there exists a ROI which the curve passes across but p does not. As ROIs are compact sets, V is covered by a union of finite number of open balls. Additionally, as $\pi(p)$ is a continuous, every open ball has interaction with at least one other ball. Then, these facts and path-connectedness of open ball give existence of paths composed of boundary points.

Let ∂V be a boundary of V . Here, assuming there exists a lifted Jordan curve q such that $\pi(q) \not\subset \partial V$ and the area of inner region surrounded by $\pi(q)$ is greater than any area of that bounded by $\pi(p)$ ($p \in [p]$), it brings contradiction. Denote one certain point in $\pi(q)$ but not in ∂V as x . As x is in the interior of V , there exist an open ball $(x \in) B \subset \text{Int}(V)$. As x is on $\pi(q)$, B is not a subset of inner region surrounded by $\pi(q)$. As an open ball is path-connected, there exist a lifted Jordan curve q' such that the region surrounded by $\pi(q')$ contains the region bounded by $\pi(q)$. Because $\pi(q')$ is a subset of V , q' is equivalence with p . Area of ball is non-zero and finite division does not make area zero. Because there exists the boundary path and non-boundary paths are not upper bounds of area of inner region bounded by it, the boundary path is an upper bound.

In the lower bound case, there is two cases. One is a case that some ROI is subsets $\text{Int}(V)$ and another is $\text{Int}(V)$ is not a superset of any ROI. In the former case, the same strategy for the proof of upper bound can be applied. The assumption that there exists a non-boundary path such that the area of inner region bounded by it is lower bound of those bounded by $\pi(p)$ ($p \in [p]$) contradicts similarly to the upper bound case. In latter cases, lower bound is clearly zero because I can make inner area of a curve as small as possible by using open ball such that is divided into two parts by the curve. $\square$

My mathematical analysis also predicts that the stride size does not affect the size-invariant responses, because my proof for the size invariance does not involve conditions for the size of intersections of ROI pairs.

3.7 Ablation of max pooling

Sections 3.5 and 3.6 demonstrated the impact of max pooling on the size-invariant responses of the first layer units to lower SFs. To further illustrate how these effects influence the acquisition of size-invariant representation in the final processing layer, I conducted an ablation study. I developed 20 models ablated max pooling layers from the SNN and trained them with the KDEP and RaFD databases. The ablation models performed modest correct rates above chance level for all facial expressions comparably to the original SNNs (0.49 ± 0.05 ; $p = 0.13$, t -test). The performance order for the seven expressions was largely consistent with the original SNNs: the performance was best for surprised (0.66) and happy (0.63) faces, moderate for angry (0.48), disguised (0.45), fearful (0.42), sad (0.40) and neutral (0.42) faces (Fig. 3.7.1A).

Examining the SF representation in the FC1 layer, we discovered a distinction from the originals (Fig. 3.3.1B, left). In the ablation model's tested units, there was an absence of object-based encoding in the low SF range (Fig. 3.7.1B). While units sensitive to low SFs below 2 cycles per object were identified, the resulting distribution exhibited a much less prominent peak compared to that of the original SNNs (Fig. 3.7.1C).

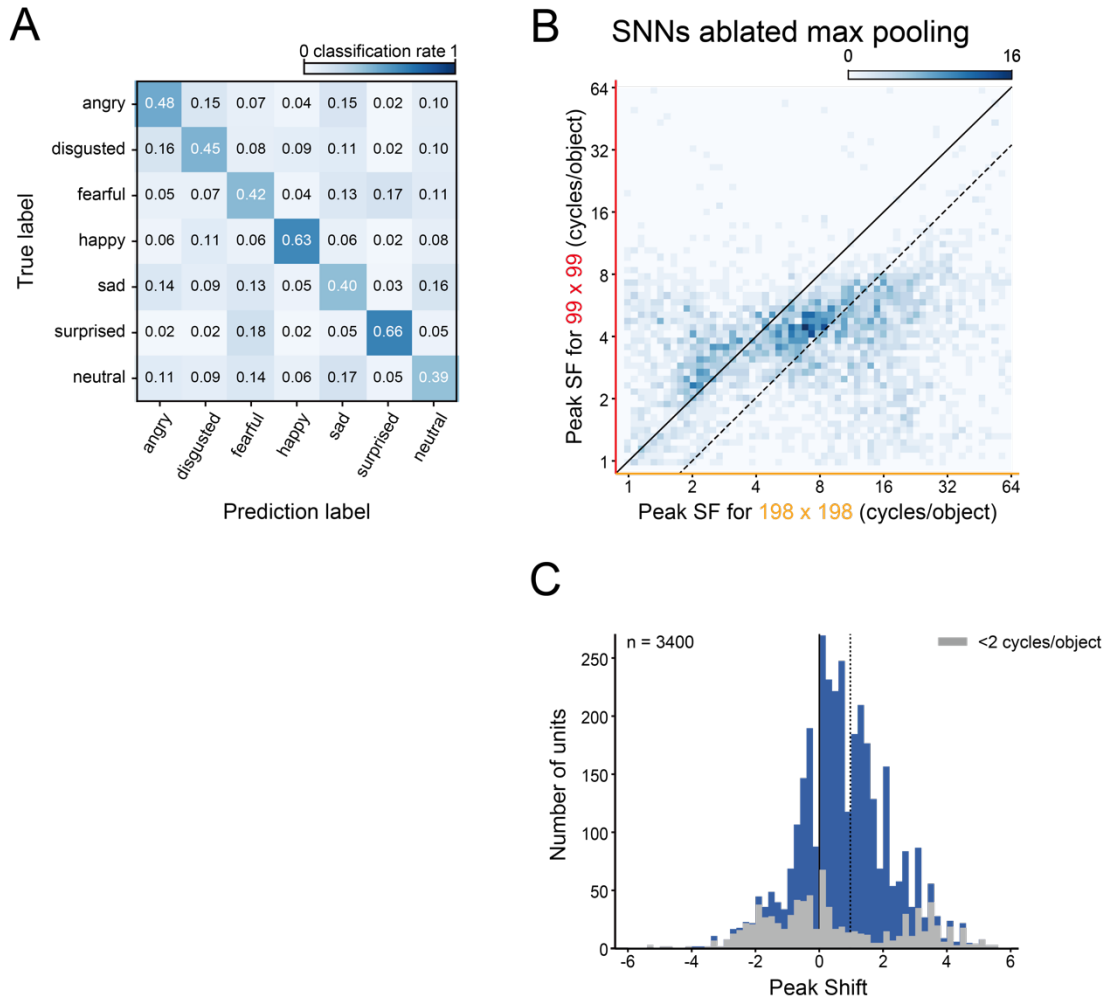


Figure 3.7.1. Effects of ablation of max pooling layers from SNNs on the classification performance of 20 the ablation model (A) and the SF tuning reference frame of FC1 units in these networks (B and C). The average performance across the model was obtained in the same way as in Fig. 3.1.1 and responses of FC1 units were obtained in the same way as in Fig. 3.3.1. (A) A confusion matrix illustrates the rate of classification of each facial expression (true label) as one of the seven expressions (prediction label). (B) A two-dimensional histogram of peak SFs obtained with large (198×198) versus small (99×99) stimulus images. (C) Distribution of peak shifts of units. Arrowheads indicate the estimated locations of multiple peaks in the distribution.

Gray columns indicate units with a response peak at SFs below 2 cycles/object for large and/or small stimuli.

These results strongly imply that max pooling plays a crucial role in constructing SF representations in the FC1 layer, particularly with respect to object-based SF representations in the lower SF range.

3.8 Effects of sliding stride on SF tuning

Finally, I examined the effect of another free parameter of our models, the stride size, on the SF sensitivity of FC1 units. I changed the stride of the two max pooling layers from 4 to 2 in the SNNs. The stride size of 2 was also employed in the narrow-pooling model. This modified model with a smaller stride of 2 achieved a mean correct rate of 0.54, which was better than that of the SNNs (0.51) but similar to the narrow pooling models (0.55) (vs. SNN, $p = 0.0020$; vs. the narrow pooling, $p = 0.23$; t -test with Bonferroni correction). An analysis of FC1 responses to two stimulus sizes (198×198 and 99×99 pixels) revealed that the distribution of the peak shift had three peaks, located at -4.17 , -0.0107 , and 0.976 (the first and second p -values < 0.001 , the third p -value = 0.098 ; excess mass test for multimodality; solid and open arrowheads in Fig. 3.8.1). Two of them were conspicuous and located near 0 and 1 (solid triangles), and the third one at the periphery of -4.17 (open triangle) was negligibly small in height (3.2% and 3.3% of the two major peaks). Comparison with Fig. 3.3.1B, C show that the small-stride model exhibited a similar SF representation as in the SNNs, in that there were two main groups of units, one sensitive to low SFs, representing SFs in the object-based reference frame (peak shift around 0), and the other sensitive to high SFs, representing SFs in the retina-based reference frame (peak shift around 1).

The change of the stride from 4 to 2 had little effects on the reference frame of SF sensitivity of FC1 units.

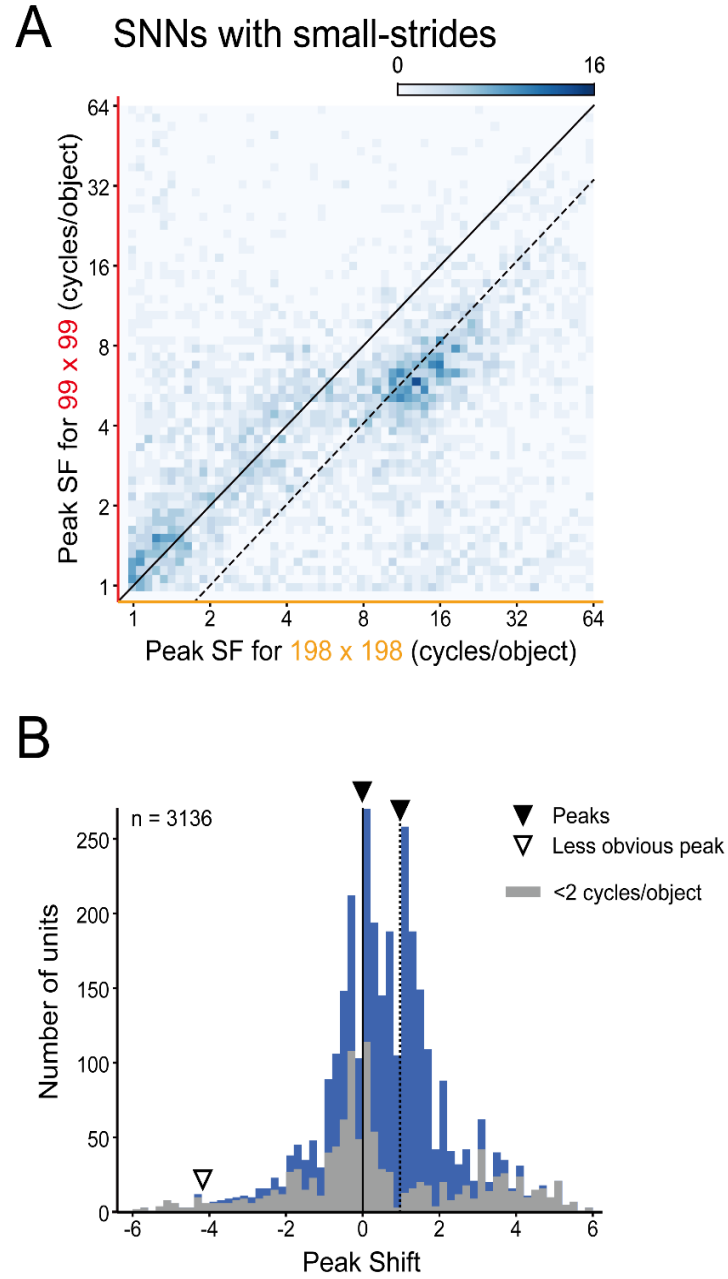


Figure 3.8.1. Effects of alternation of sliding strides of pooling windows on SF tuning reference frames of FC1 units of the SNNs. The responses of the FC1 units were obtained in the same way as in Fig. 3.3.1. (A) A two-dimensional histogram of peak SFs obtained with large (198×198)

versus small (99×99) stimulus images. (B) Distribution of peak shifts of units. Arrowheads indicate the estimated locations of multiple peaks in the distribution.

Discussion

This study presents an initial endeavor to construct and investigate a computational model for facial expression processing along the subcortical pathway. While DNNs have achieved notable accomplishments in modeling visual processing in the ventral cortical pathway, it has remained unclear whether and how the CNN architecture can be adapted to processing in the subcortical pathway. Here I demonstrated that the SNNs, which implemented the three subcortical properties (shallow processing, DoG-type filters, and wide-area spatial pooling; Fig. 2.1.1), can be trained to classify facial expressions. Compared to DNNs previously tested for facial expression classification, the SNNs exhibited modest classification accuracy, akin to individuals with affective blindness (Fig. 3.1.1). Replacement of any one of these properties with the corresponding cortical properties resulted in higher performances (Fig. 3.2.1). Replacement of a combination of two or three properties further improved classification performances in a partially additive manner (Fig. 3.2.1). Substituting any of these properties with their cortical counterparts improved performance, indicating the significance of all three subcortical properties as factors that limit performance of the SNNs. In the FC1 layer, units were divided into two groups: one sensitive to retina-based SFs and the other sensitive to object-based SFs. The number of retina-based units was reduced in all but one of the modified models, suggesting that the three features are also important for preserving retina-based SF representation (Figs. 3.3.1, 3.3.2). Max pooling over the wide window employed in the SNNs contributed to object-based SF representation by units sensitive to low SFs (Fig. 3.5.1). These findings suggest that the three processing properties of the subcortical pathway have a limiting effect on facial expression recognition. These insights shed light on the computational processes employed by individuals with affective blindness and newborns.

4.1 Modest performance of the SNNs and roles of the subcortical pathway properties

Based on psychological assessment and brain imaging in V1-lesioned patients, it has been proposed that affective blindsight is mediated by components of the subcortical pathway spared by the lesions, including the superior colliculus, pulvinar, and amygdala (de Gelder et al., 1999; Pegna et al., 2005; Striemer et al., 2019). One view assumes that the shortest route directly connecting the three subcortical structures conveys facial expression information from the superior colliculus via the pulvinar to the amygdala (Tamietto and de Gelder, 2010). A different view proposes that information from the pulvinar reaches the amygdala through the facial processing system in the temporal cortex, under an assumption that “the direct connections of the pulvinar with the amygdala are likely insufficient in themselves for recognizing emotional expressions” (Gerbella et al., 2019). The present study demonstrates that the SNNs with only three processing stages and the subcortical physiological properties can successfully acquire an ability to discriminate facial expressions.

The average correct rate (0.49–0.51) of classifying the seven facial expressions in present study was well above chance ($1/7 = 0.14$), but was far from perfect. The modest correct rate is in line with the performance of patients with affective blindsight. Pegna et al. (2005) reported that a patient with bilateral lesions in V1 discriminated happy faces from either angry, sad, or horrified faces at correct rates of 0.58–0.62, only marginally above the chance level of 0.5. Another patient with bilateral lesions in V1 exhibited correct rates of 0.64–0.67 for happy vs. fearful or angry faces (chance level = 0.5; Striemer et al. 2019). The residual ability of facial expression classification in these patients was only moderate compared to the nearly perfect performance in

healthy people. This raises the question of why subcortical processing supports vision more poorly than visual functions mediated by the cortical pathway.

A traditional explanation is that neurons in the subcortical pathway respond to low SFs and are less sensitive to high SFs than the cortical pathway: 0.56–11.11 cycles/degree for the superior colliculus (Chen et al., 2018), <1.2–1.4 cycles/degree for the pulvinar (Van Le et al., 2013), and <3.24 cycles/degree for the amygdala (Vuilleumier et al., 2003; Méndez-Bértolo et al., 2016). This predisposition towards low SF sensitivity will limit the ability of the subcortical pathway to analyze the fine details of visual images and can itself result in the inaccurate processing of face images. However, some researchers dispute the dependence of the subcortical response on low SFs (De Cesare and Codispoti, 2013; McFadyen et al., 2017). Our results show that low SF sensitivity, if important, was not the only cause because the DoG filter models combined with narrow-pooling or add-layer modifications exhibited improved performances, despite the fact that our DoG filters were tuned to low SFs, with a preferred SF range between 0.17 and 3.4 cycles/degree (see Appendix A1 for more details). Note that we estimated this value on an assumption of an image size of 30.5° based on our DoG parameters, the filter resolution, and the RF size of superior colliculus neurons representing the foveal region. The range of DoG-filter width is slightly wider than that applied in a recent model of the superior colliculus (0.36-1.88 cycles/degree based on our calculation; Méndez et al., 2022).

Another explanation is that the small number of processing stages in the cortical pathway hampers detailed analysis of visual inputs. However, a previous study (Dailey et al., 2002) showed that CNNs that had only two processing layers, with Gabor filters at the first stage, performed highly accurate discrimination of facial expressions (the mean correct rate for classifying six facial

expressions was 0.90). The performance of my SNNs incorporating the three subcortical properties was not this high. This was not due to inadequate training, because the performance reached a plateau and stayed stable over a large number of iterations in the training sessions (Fig. 3.1.1). This was verified further by showing that the correct performance for the test set remained unchanged even after overly excessive training with 3,000,000 iterations. Furthermore, replacing not only the small number of processing layers but also the filter type at the first processing layer and the width of the pooling window with the corresponding cortical properties improved the performance of the SNNs (Fig. 3.2.1). These three properties at least partially underlie the less detailed processing of facial images in the subcortical pathway, and hence, may be responsible for the low performance in affective blindsight.

4.2 Confusions of facial expressions in the SNNs, DNNs and patients

The accuracy of classification by the SNNs varied across facial expressions (Fig. 3.1.1A, B). The classification performance of the SNNs was best for happy and surprised faces and worst for sad and neutral faces. The rank order of performance on the seven facial expressions was consistent across the 20 SNNs trained independently from random states (Fig. 3.1.1C). It also corresponds to the classification performance by previously developed AlexNet-based DNNs (Inagaki et al., 2022a). These DNNs were trained to discriminate between the seven expressions derived either from the KDEF database or the Kokoro Research Center (KRC) facial expression database (Ueda et al., 2019). Like the SNNs, the DNNs exhibited the best performance for happy and surprised faces (KDEF: 0.93 for happy, 0.84 for surprised; KRC: 0.91 for happy, 0.84 for surprised; chance level, 0.14). This coincidence may simply suggest that within each database, facial features are consistent across faces with happy or surprised expressions, but are more diverse across faces with sad or neutral expressions. However, the variations across examples of facial expressions

within a database are not the sole reason for the difference in the performance across facial expressions, because neutral faces were classified poorly by the SNNs (Fig. 3.1.1B; correct rate = 0.34), but the DNNs of Inagaki et al. (2022) classified them with high correct rates (0.85 for KDEF, 0.82 for KRC). An alternative, yet-to-be-tested explanation is that the ease (or difficulty) of classification may vary across the facial expressions owing to differences in conspicuousness of component actions underlying different expressions. Different neural networks may have shown partly similar expression-specific performance according to these differences.

The relatively poor ability to distinguish between sad and neutral faces was also observed in another CNN with the first layer of DoG filtering and average pooling (Méndez et al., 2022). This CNN was constructed for simulating facial processing in the superior colliculus, and was trained to discriminate three facial expressions: happy, sad, and neutral. The CNN showed the best performance for happy faces and moderate performance for sad faces, but classified neutral faces into neutral faces with a classification rate of 0.49 and into sad faces at 0.39. The fact that this CNN and the SNNs in the present study demonstrated this confusion, whereas AlexNet-based DNNs and our add-layer models (average performance, 0.68; 0.52 for sad, 0.67 for neutral) did not, suggests that the convolution processes after the initial DoG filtering (in the case of add-layer models) or the convolution by the Gabor filters (in the case of AlexNet-based DNNs) may be critical for classification of sad and neutral faces.

The expression-dependent performance of the SNNs also had both similarities and dissimilarities to that observed in a V1-lesioned patient. The patient reported by de Gelder et al. (1999) classified happy and sad faces with a higher correct rate than angry and fearful faces; our SNNs and this

patient classified happy faces well, whereas the performance for sad faces was poor in the SNNs but good in the patient.

Our SNNs were solely trained using facial images under controlled laboratory conditions. A crucial area for future research lies in assessing the performance of SNNs in expression classification of facial images taken under more natural and diverse conditions (Li and Deng, 2022). Additionally, we note that although we used human face photographs as training and test images, our neural network models were constructed based on physiological findings from both monkeys and humans. This raises the question of how effectively the SNNs trained with human faces perform in classifying monkey facial expressions in cross-species scenarios. It should be acknowledged, however, that the facial expressions in these two species differ, and there is ongoing controversy regarding the interpretation of facial expressions in nonhuman primates (e.g., see Sterck and Goossens, 2008; Beisner and McCowan, 2014; Waller et al., 2020; Taubert and Japee, 2021). Furthermore, it is worth noting that the performance of deep neural networks does not necessarily generalize completely even for human faces from different countries (Inagaki et al., 2022b).

4.3 Reference frame of coding SF information

FC1 units of the SNNs consisted of two major groups, each with different properties regarding SF processing (Fig. 3.3.1). One group of units responded best to the same object-based SFs (cycles/object) regardless of the stimulus size (peak shift around 0). This size-invariant response indicates that these units represent SFs in the object-based coordinate. Most of these units were tuned to a low SF range (around 1–2 cycles/object). In the other group of units, the optimal object-based SFs shifted when testing was performed with different stimulus sizes. The direction of the

shift was consistent with the interpretation that the units were tuned to retina-based SFs (cycles/degree) (peak shift around 1). That is, for larger stimuli, the units responded to higher object-based SFs that corresponded to the same retina-based SFs. The DoG filters at the initial stage and the shallow architecture appear to be critical for preserving the SF representation based on the retina-based coordinate, because FC1 units with retina-based SF sensitivity were reduced in number when I changed the filters in the first convolution layer to Gabor filters or increased the number of processing layers (Fig. 3.3.2A, middle, bottom).

While most of the object-based SF units of the SNNs were tuned to low SFs, the majority that were retina-based were tuned to high SFs (Fig. 3.3.1B). This curious dichotomy likely resulted from the wide max pooling process. The max pooling operation renders units in the pooling layer downstream of the same region of interest of the convolution layer insensitive to retina-based SFs. In networks, retina-based SFs are encoded in positional information by convolution of DoG or Gabor filters. The collapse of local positional information by the max pooling impairs the information about retina-based SFs, making it difficult for later units, e.g., FC1 units, to create object-based SF sensitivity by integrating an appropriate range of retina-based SFs as do inferior temporal cortex neurons in the monkey (Inagaki and Fujita, 2011). However, the max pooling operation can provide another way of supporting size-invariant responses to face images in a low SF range. Low pass-filtered facial images contain only simple shapes such as solid circles at eye or mouth positions. DoG or Gabor filters can extract the edges of simple shapes, and the max pooling blurs the positional information, allowing low pass-filtered images of different sizes to elicit the same responses. When the pooling window is narrower, this effect would be incomplete, and units with intermediate peak shift values would increase, as I found in the narrow-pooling models (Fig. 3.3.2A, top). My mathematical analysis provided the fact that the strength of this

size-invariant effect by the max pooling operation depends on the width of ROIs of max pooling. How large change are ignored by the pooling characterized by area of ∂V . Though area of ∂V depend on how to take ROIs but not on how great overlapping of ROIs. This result predicts the experimental results that SNN with small-strides have similar SF representation with SNN.

One may wonder why FC1 units of the SNNs maintain sensitivity to retina-based SFs, i.e., size-dependent representation of SFs, despite the demand that I imposed on the SNNs to classify facial expressions regardless of the seven different face image sizes. One possibility is that the architecture of our SNNs cannot achieve sufficient object-based representation, and remains suboptimal for the required task even after the excessive training sessions. This may be a reason for the modest classification performance of the SNNs. Indeed, replacement of the subcortical processing properties with the cortical properties resulted in the representation becoming more object-based (Fig. 3.3.1B, C, Fig. 3.3.2) and improved the classification performance (Fig. 3.2.1). If the optimal performance under our training conditions requires a high degree of object-based SF encoding, the models that showed object-based SF encoding should show the highest correct rate, but this was not the case. The add-layer models and the narrow-pooling + add-layer models exhibited the clearest object-based encoding of SFs (Figs. 3.3.2A bottom, 3.3.2B middle), while they performed worse than the full-replacement model (Figs. 3.2.1A, 3.3.1C).

In the primate amygdala, the responses of many neurons are affected by retina-based SFs, and neurons responding to object-based SFs are a minority (Inagaki and Fujita, 2010). By contrast, many FC1 units tuned to low SFs of the SNNs exhibit object-based SF sensitivity. The paucity of evidence for object-based SFs in the amygdala may be related to the fact that the previous

electrophysiological study (Inagaki and Fujita, 2010) did not present face images of very low SFs in their experiments, and may have overlooked the neurons with object-based SF sensitivity.

Some inferior temporal cortex neurons have an invariance in their responses to size and homeomorphic change of shapes (Rolls and Baylis, 1986; Ito et al., 1995; Komatsu and Ideura, 1993). The max pooling operation may help achieve the invariant responses. The max pooling ignores some degree of positional changes of inputs in each region of interest. Because size and homeomorphic changes involve changes of edge positions without changing topologies, if the changes are small enough to be covered by each region of interest, stimuli before and after the change would yield similar responses. The results shown in Figure 3.4.1 and my mathematical proof suggest that some aspects of the invariant responses of inferior temporal cortex neurons can be simply formed by bypassing lower visual cortical areas such as V1. Such shortcut routes indeed exist, including the projection from the pulvinar to V2 then to posterior inferior temporal cortex and the projection from the pulvinar to V4 then to anterior inferior temporal cortex (Pessoa and Adolphs, 2010).

Conclusion

We have presented an original computational approach to investigate the facial expression processing in the subcortical pathway of primates. This approach analyzes the performance of SNNs with subcortical properties and the effect of replacing these properties with corresponding cortical counterparts. We have demonstrated that the performance of the SNNs is constrained by the three key subcortical properties: the shallowness of processing stages, the DoG-type receptive fields at the initial stage, and spatial pooling over a wider visual field. These properties also affect the reference frame of spatial frequency representation in the final layer of the SNNs. Our findings offer insights into the neural mechanisms and functions of the subcortical pathway, highlighting their distinction from those of the cortical pathway. They provide an explanation for the limited proficiency observed in facial expression recognition among individuals with impaired or undeveloped visual cortices.

An important direction for future research is to experimentally assess the validity of our SNNs as proxies for biological neural network models for the subcortical pathway. One possible approach is to generate facial expression images that elicit the strongest responses of units in either the SNNs or the full-replacement models using activation maximization procedures (Nguyen et al., 2016; Bashivan et al., 2019). These two types of images can then be used to evaluate the performance of human observers in classifying the facial expressions. Further, neural responses can be measured by single-neuron recordings in non-human primates or functional magnetic resonance imaging in human observers. Employing these types of approaches would enable us to evaluate the extent to which the SNNs are linked to biological neural systems.

Research interest in the role of subcortical structures in cognitive functions has recently surged, but physiological data are still much sparser for subcortical structures than for the cerebral cortex (Janacsek et al., 2022). Computational approaches, such as the one we have presented here, are expected to partially compensate for this data scarcity and to guide future research.

Appendix

A1 SF tunings of DoG filters

I estimated what spatial frequencies DoG filters tune by using full width of half maximum (FWHM) of DoG function on the spatial frequency domain. A FWHM of function f is the difference between two argument values x_1, x_2 at which images of them under f are equal to half of the maximum value on codomain. Given Fourier transformed DoG F , the points taking the half maximum are calculated by solving $F(x) = A/2$, where A is the maximum. In the following section, first I calculate the Fourier transformed DoG F . After that, I solve the equation and estimate FWHMs of DoG filters with numerical solutions.

The 2D Difference of Gaussians is defined in $\mathbb{R}^2$ as

$$f(x, y; k_1, k_2, \sigma_1, \sigma_2) = \frac{k_1}{\sigma_1} \exp\left(-\frac{x^2 + y^2}{2\sigma_1^2}\right) - \frac{k_2}{\sigma_2} \exp\left(-\frac{x^2 + y^2}{2\sigma_2^2}\right), \quad (1)$$

where $k_1, k_2, \sigma_1, \sigma_2$ are nonnegative real values and $\sigma_1 > \sigma_2$. As it is a symmetric function on x, y , the two variables can be reduced to the nonnegative real valuable r as the following:

$$f(r) = \frac{k_1}{\sigma_1} \exp\left(-\frac{r^2}{2\sigma_1^2}\right) - \frac{k_2}{\sigma_2} \exp\left(-\frac{r^2}{2\sigma_2^2}\right) \quad (2)$$

Adapted the Fourier transformation to eq. 2, I got DoG on Fourier domain expressed as

$$F(\omega) = \sqrt{2}k_1 \exp\left(-\frac{\sigma_1^2 \omega^2}{2}\right) - \sqrt{2}k_2 \exp\left(-\frac{\sigma_2^2 \omega^2}{2}\right), \quad (3)$$

where ω is the angular frequency. the equation $dF/d\omega = 0$ gives the critical point ω_p . Designating

$$K = \sqrt{(k_2/k_1)}, S = \sigma_2/\sigma_1,$$

$$\omega_p = \frac{2}{\sigma_1} \sqrt{\frac{\log KS}{S^2 - 1}}. \quad (3)$$

Then, the equation for obtaining FWHM is obtained as the solution of the following equation:

$$K^2 x^{S^2} - x + C = 0 \quad C = \frac{1}{2} (KS)^{-\frac{2}{S^2-1}} \left(1 - \frac{1}{S^2}\right) \quad (4)$$

where x is a designation of $\exp(-\omega^2 \sigma^2/2)$. When eq. 4 has two solutions $x_1 < x_2$, FWHM is as below.

$$FWHM = \frac{\sqrt{-2 \log x_1}}{2\pi\sigma_1} - \frac{\sqrt{-2 \log x_2}}{2\pi\sigma_1} \quad (5)$$

When eq. 4 has only one solution x_1 , it means $F(0) > F(0)/2$. Theus,

$$FWHM = \frac{\sqrt{-2 \log x_1}}{2\pi\sigma_1} \quad (6)$$

For the case of $A_2 = 0$, FWHM is clearly $\sqrt{(2\log 2)/(2\pi\sigma_1)}$. While it seems hard to solve Eq. 4 analytically, a numerical method is applicable. the solutions are calculated with Wolfram Alpha (Wolfram Research Inc., 2023). As my DoG filters is composed of 11×11 pixels and the smallest receptive field of SC neurons is 1.5° , one degree corresponds to 7.4 pixels. Then, given the lower

and upper ones of half-magnitude points, frequencies with cycles/degree are calculated by multiplying them and 7.4/11.

Table 1. FWHMs of DoG filters. f_p represents peak frequency. f_l and f_h stand for lower and upper ones of half-magnitude points. the frequency unit is cycles per degree.

σ_1	$K^2 = k_2/k_1$	$S = \sigma_2/\sigma_1$	f_p	$FWHM(f_l, f_u)$
$1/2\sqrt{2}$	—	—	0	0.43
$1/4\sqrt{2}$	—	—	0	0.87
$1/8\sqrt{2}$	—	—	0	1.68
$1/16\sqrt{2}$	—	—	0	3.4
$1/2\sqrt{2}$	8/5	2	0.337	0.322(0.217, 0.540)
$1/4\sqrt{2}$	8/5	2	0.674	0.643(0.436, 1.078)
$1/2\sqrt{2}$	32/5	4	0.238	0.272(0.170, 0.442)
$1/4\sqrt{2}$	32/5	4	0.476	0.544(0.340, 0.885)
$1/2\sqrt{2}$	2	2	0.338	0.305(0.247, 0.552)
$1/4\sqrt{2}$	2	2	0.713	0.610(0.494, 1.104)
$1/2\sqrt{2}$	8	4	0.244	0.267(0.178, 0.445)
$1/4\sqrt{2}$	8	4	0.487	0.534(0.357, 0.891)
$1/2\sqrt{2}$	12/5	2	0.372	0.293(0.268, 0.561)
$1/4\sqrt{2}$	12/5	2	0.744	0.587(0.537, 1.123)
$1/2\sqrt{2}$	48/5	4	0.248	0.264(0.184, 0.448)
$1/4\sqrt{2}$	48/5	4	0.496	0.527(0.369, 0.895)

Acknowledgement

First and foremost, I am deeply grateful to Drs. Mikio Inagaki and Ichiro Fujita for continuous support, helpful advice, and patience. They gave an opportunity to start this project and continuously encouraged me to proceed with this project. Their constant trust, gentle prodding, immense knowledge, and plentiful experience have guided me in all the time of my academic research and daily life. I must also be grateful to Dr. Takashi Shinozaki for his invaluable advice on from a fundamental layer to a practical layer of this project. Without his helpful support and critical comments, I could not have accomplished this project.

I would like to offer my special thanks to Dr. Shinji Nishimoto for his accepting me into his group and, furthermore, serving as a supervisor. I would like to express my appreciation to Drs. Shigeru Kitazawa, Takeshi Yagi, and Hiroshi Tamura for serving the dissertation committee.

I thank all of my prior and current members of my laboratory. They gave me kind advice and helped me with my daily things. Also, I thank to Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science.

Finally, I would like to thank my family. They gave me nothing but support both emotionally and financially throughout the period of my Ph.D. course. My gratitude for them can hardly be expressed in words.

References

- Afraz, S.R., Kiani, R. and Esteky, H. Microstimulation of inferotemporal cortex influences face categorization. *Nature*, **442**, pp.692-695 2006.
- Amaral, D.G. and Price, J.L. Amygdalo-cortical projections in the monkey (*Macaca fascicularis*). *J. Comp. Neurol.*, **230**, pp.465-496 (1984).
- Ameijeiras-Alonso, J., Crujeiras, R.M. and Rodríguez-Casal, A. Mode testing, critical bandwidth and excess mass. *Test*, 28, pp.900-919 (2019).
- Ameijeiras-Alonso, J., Crujeiras, R.M. and Rodríguez-Casal, A. Multimode: An R package for mode assessment. *J. Stat. Softw.* **97**, Issue 9 (2021).
- Bashivan, P., Kar, K. & DiCarlo, J.J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019). doi: 10.1126/science.aav9436
- Beisner, B.A. & McCowan, B. Signaling context modulates social function of silent bared - teeth displays in rhesus macaques (*Macaca mulatta*). *Am. J. Primatol.* 76, 111-121 (2014). doi: 10.1002/ajp.22214
- Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000). <https://opencv.org/> (version, 2.4.8; this version is no longer available)
- Buiatti, M., Di Giorgio, E., Piazza, M., Polloni, C., Menna, G., Taddei, F., Baldo, E. & Vallortigara, G. Cortical route for facelike pattern processing in human newborns. *Proc. Natl. Acad. Sci. USA*. **116**, 4625–4630 (2019).
- Cassia, V.M., Simion, F. & Umiltà, C. Face preference at birth: the role of an orienting mechanism. *Develop. Sci.* **4**, 101-108 (2001).
- Chen, C.-Y., Hoffmann, K.-P., Distler, C., & Hafed, Z. M. The foveal visual representation of the primate superior colliculus. *Curr. Biol.* **29**, 2109-2119 (2019).

- Chen, C.Y., Sonnenberg, L., Weller, S., Witschel, T. & Hafed, Z.M.. Spatial frequency sensitivity in macaque midbrain. *Nat. Comm.*, 9, pp.1-13 (2018).
- Cheng, K., Saleem, K.S., Tanaka, K. Organization of corticostriatal and corticoamygdalar projections arising from the anterior inferotemporal area TE of the macaque monkey: a Phaseolus vulgaris leucoagglutinin study. *J Neurosci* 17:7902-7925 (1997).
- Churan, C.-Y., Guitton, D. & Pack, C.C. Spatiotemporal structure of visual receptive fields in macaque superior colliculus. *J. Neurophysiol.* **108**, 2653-2667 (2012).
- Connor, C.E., Brincat, S.L. & Pasupathy, A. Transformation of shape information in the ventral pathway. *Curr. Opin. Neurobiol.* **17**, 140-147 (2007).
- Conway, B.R., Chatterjee, S., Field, G.D., Horwitz, G.D., Johnson, E.N., Koida, K. & Mancuso, K. Advances in color science: from retina to behavior. *J. Neurosci.* **30**, 14955-14963 (2010).
- Cynader, M. & Berman, N. Receptive-field organization of monkey superior colliculus. *J. Neurophysiol.* **35**, 187-201 (1972).
- Dailey, N.M., Cottrell, W.G., Padgett, C. & Adolphs, R. EMPATH: A neural network that categorizes facial expressions. *J Cogni Neurosci.* **14**, 1158-1173 (2002).
- de Gelder, B., Vroomen, J., Pourtois, G. & Weiskrantz, L. Non-conscious recognition of affect in the absence of striate cortex. *NeuroRep.* **10**, 3759-3763 (1999).
- Duchaine, B. & Yovel, G. A revised neural framework for face processing. *Annu. Rev. Vis. Sci.* **1**, 393-416 (2015).
- Elorette, C., Forcelli, P.A., Saunders, R.C. & Malkova, L. Colocalization of tectal inputs with amygdala-projecting neurons in the macaque pulvinar. *Front Neural Circuits.* **12**, 91; 10.3389/fncir.2018.00091 (2018).
- Freeman, J. & Simoncelli, E.P. Metamers of the ventral stream. *Nat. Neurosci.* **14**, 1195-1201 (2011).

- Freiwald, W., Duchaine, B. & Yovel, G. Face processing systems: from neurons to real-world social perception. *Annu. Rev. Neurosci.* **39**, 325-346 (2016).
- Fukushima, K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193-202 (1980).
- Gerbella, M., Caruana, F. & Rizzolatti, G. Pathways for smiling, disgust and fear recognition in blindsight patients. *Neuropsychologica.* **128**, 6-13 (2019).
- Goldberg, M. E. & Wurtz, R. H. Activity of superior colliculus in behaving monkey. I. Visual receptive fields of single neurons. *J. Neurophysiol.* **35**, 542–559 (1972).
- Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. Cambridge: MIT Press, pp. 71–73 (2016). <http://www.deeplearningbook.org>
- Gothard, K.M., Battaglia, F.P., Erickson, C.A., Spitler, K.M. & Amaral, D.G. Neural responses to facial expression and face identity in the monkey amygdala. *J. Neurophysiol.* **97**:1671–1683 (2007).
- Güçlü, U. & van Gerven, M.A.J. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005-10014 (2015).
- Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron.* **95**, 245-258 (2017).
- Hasselmo, M.E., Rolls, E.T. and Baylis, G.C. The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behav. Brain Res.*, **32**, pp.203-218 (1989).
- Hayes, T., Morrone, M.C., Burr, D.C. Recognition of positive and negative bandpass-filtered images. *Perception* **15**, 595–602 (1986).
- Haxby, J.V, Hoffman, E.A. & Gobbini, M.I. The distributed human neural system for face perception. *Trends Neurosci.* **4**, 223-233 (2000).

- He, K., Zhang, X., Ren, S. and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proc. IEEE Int. Conf. Comput. Vis.* pp. 1026-1034 (2015).
- He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 770-778 (2016).
- Hubel, D.H. and Wiesel, T.N. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. neurophysiol.*, **28**, pp.229-289. (1965)
- Hubel, D.H. and Wiesel, T.N. Ferrier lecture-Functional architecture of macaque monkey visual cortex. *Phil. Trans. R. Soc. B.* **198**, pp.1-59 (1977).
- Inagaki, M. & Fujita, I. Reference frames for spatial frequency in face representation differ in the temporal visual cortex and amygdala. *J. Neurosci.* **31**, 10371-1039 (2011).
- Inagaki, M., Inoue, K.-I., Tanabe, S., Kimura, K., Takada, M. & Fujita, I. Rapid processing of threatening faces in the amygdala of nonhuman primates: subcortical inputs and dural roles. *Cereb. Cortex.* 10.1093/cercor/bhac109 (2022a).
- Inagaki, M., Ito, T., Shinozaki, T. & Fujita, I. Convolutional neural networks reveal differences in action units of facial expressions between face image databases developed in different countries. *Front. Psychol.* doi.org/10.3389/fpsyg.2022.988302 (2022b).
- Ito, M., Tamura, H., Fujita, I. & Tanaka, K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* **73**, 218-226 (1995).
- Janacsek, K., Evans, T.M., Kiss, M. Shah, L., Blumenfeld, H. & Ullman, M.T. Subcortical cognition: the fruit below the rind. *Annu. Rev. Neurosci.* **45**, 361-386 (2022).
- Jones, J.P. & Palmer, L.A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1233–1258 (1987).
- Johnson, M.H. Subcortical face processing. *Nat. Rev. Neurosci.* **6**, 766-774 (2005).

- Johnson, M.H., Senju, A. & Tomalski, P. The two-process theory of face processing: modifications based on two decades of data from infants and adults. *Neurosci. Behav. Rev.* **50**, 169-179 (2015).
- Kar, K., Kubilius, J., Schmidt, K., Issa, E.B. and DiCarlo, J.J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. neurosci.*, **22**(6), pp.974-983 (2019).
- Knudsen, E.I. Evolution of neural processing for visual perception in vertebrates. *J. Comp. Neurol.* **528**, 2888-2901 (2020).
- Komatsu, H. & Goda, N. Neural mechanisms of material perception: Quest on Shitsukan. *Neurosci.* **392**, 329-347 (2018).
- Komatsu, H. and Ideura, Y. Relationships between color, shape, and pattern selectivities of neurons in the inferior temporal cortex of the monkey. *J. Neurophysiol.* **70**, pp.677-694 (1993).
- Kravitz, D.J., Saleem, K.S., Baker, C.I., Ungerleider, L.G. & Mishkin, M. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* **17**, 26-49 (2013).
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Proces. Syst (NeurIPS)*. **25**, 1097-1105 (2012).
- Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T. & Van Knippenberg, A.D. Presentation and validation of the Radboud Faces Database. *Cognition and emotion*, **24**, 1377-1388 (2010).
- Le, Q.V., Isbell, L.A., Matsumoto, J., Nguyen, M., Hori, E., Maior, R.S., Tomaz, C., Tran, A.H., Ono, T. and Nishijo, H. Pulvinar neurons reveal neurobiological evidence of past selection

- for rapid detection of snakes. *Proceedings of the National Academy of Sciences*, **110**, 19000-19005 (2013).
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature*. **521**, 436-444 (2015).
- LeDoux, J.E. Emotion circuits in the brain. *Annu. Rev. Neurosci.* **23**, 155–184 (2000).
- Leonard, C.M., Rolls, E.T., Wilson, F.A.W. & Baylis, G.C. Neurons in the amygdala of the monkey with responses selective for faces. *Behav. Brain Res.* **15**, 159–176 (1985).
- Li S. & Deng, W. Deep facial expression recognition: a survey. *IEEE Trans. Affect. Comput.* **13**, 1195-1215. doi: 10.1109/TAFFC.2020.2981446 (2022).
- Lundqvist, D., Flykt, A. & Öhman, A. The Karolinska directed emotional faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institute, 91, 630 (1998).
- Marino, R.A., Rodgers, C.K., Levy, R. & Munoz, D.P. Spatial relationships of visuomotor transformations in the superior colliculus map. *J. Neurophysiol.* **100**, 2564-2576 (2008).
- Méndez, C.A., Celeghin, A., Diano, M., Orsenigo, D., Ocak, B. & Tamietto, M. A deep neural network model of the primate superior colliculus for emotion recognition. *Phil. Trans. R. Soc. B* **377**, 20210512 (2022).
- Méndez-Bértolo, C., Moratti, S., Toledano, R., Lopez-Sosa, F., Martinez-Alvarez, R., Mah, Y.H., Vuilleumier, P., Gil-Nagel, A. & Strange, B.A. A fast pathway for fear in human amygdala. *Nature Neurosci.*, **19**, 1041-1049 (2016).
- Morawetz, C., Baudewig, J., Treue, S. & Dechent, P. Diverting attention suppresses human amygdala responses to faces. *Front. Human Neurosci.* **4**, 226 (2010).
- Morris, J.S., DeGelder, B., Weiskrantz, L. & Dolan, R.J. Differential extrageniculostriate and amygdala responses to presentation of emotional faces in a cortically blind field. *Brain*. **126**, 1241-1252 (2001).

- Morris, J.S., Öhman, A. & Dolan, R.J. A subcortical pathway to the right amygdala mediating “unseen” fear. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1680-1685 (1999).
- Movellan, J.R.. Tutorial on Gabor filters. *Open Source Document*. **40**, 1–23 (2008).
<https://inc.ucsd.edu/mplab/75/media//gabor.pdf> (access, 2022/12/31; doi is not available)
- Nakamura, K., Mikami, A. & Kubota K. Activity of single neurons in the monkey amygdala during performance of a visual discrimination task. *J. Neurophysiol.* **67**, 1447–1463 (1992).
- Nakano, T., Higashida, N. & Kitazawa, S. Facilitation of face recognition through the retino-tectal pathway. *Neuropsychol.* **51**, 2043-2049 (2013).
- Näsänen, R. Spatial frequency bandwidth used in the recognition of facial images. *Vision Res.* **39**, 3824-3833 (1999).
- Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T. & Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Adv. Neural Inf. Proces. Syst* (NeurIPS). **29** (2016).
- Nguyen, M.N., Hori, E., Matsumoto, J., Tran, H.T., Ono, T. & Nishijo, H. Neuronal responses to face-like stimuli in the monkey pulvinar. *Eur J Neurosci.* **37**, 35-51, (2013).
- Nguyen, M.N., Matsumoto, J., Hori, E., Maior, R.S., Tomaz, C., Tran, A.H., Ono, T., & Nishijo, H. Neuronal responses to face-like and facial stimuli in the monkey superior colliculus. *Front. Behav. Neurosci.* **8**, 85 (2014).
- Pegna, A.J., Khateb, A., Lazeyras, F., & Seghier, M.L. Discriminating emotional faces without primary visual cortices involves the right amygdala. *Nat. Neurosci.* **8**, 24-25 (2005).
- Pessoa, L. & Adolphs, R. Emotion processing and the amygdala: from a “low road” to “many roads” of evaluating biological significance. *Nat. Rev. Neurosci.* **11**, 773-782 (2010).
- Petry, H.H. & Bickford, M.E. The second visual system of the tree shrew. *J. Comp. Neurol.* **527**, 679-693 (2019).

- Rai, M. & Rivas, P. A review of convolutional neural networks and Gabor filters in object recognition. *2020 Int. Conf. Comput. Sci. Comput. Intelligence (CSCI)* 1560-1567 (2020)
- Rizzolatti, G., Buchtel, H.A., Camarda, R., & Scandolara, C. Neurons with complex visual properties in the superior colliculus of the macaque monkey. *Exp. Brain Res.* **38**, 37-42 (1980).
- Roe, A.W., Chelazzi, L., Connor, C.E., Conway, B.R., Fujita, I., Gallant, J.L., Lu, H., & Vanduffel, W. Toward a unified theory of visual area V4. *Neuron* **74**, 12-29 (2012).
- Rolls E.T., Baylis G.C. Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey. *Exp Brain Res.* **65**, 38–48 (1986).
- Rutishauser, U., Tudusciuc, O., Neumann, D., Mamelak, A.N., Heller, C., Ross, I., Phiopott, L., Sutherling, W.W. & Adolphs, R.. Single-unit responses selective for whole faces in the human amygdala. *Curr. Biol.* **21**, 1654–1660 (2011).
- Sadagopan, S., Zarco, W. and Freiwald, W.A. A causal relationship between face-patch activity and face-detection behavior. *Elife*, **6**, p.e18558 (2017).
- Schmolesky, M.T., Wang, Y, Hanes, D.P., Thompson, K.G., Leutgeb, S., Schall, J.D. & Leventhal, A.G. Signal timing across the macaque visual system. *J. Neurophysiol.* **79**, 3272-3278 (1998).
- Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv* 1409.1556 (2014).
- Sowden, P.T. & Schyns, P.G. Channel surfing in the visual brain. *Trends Cogn. Sci.* **10**, 538-545 (2006).
- Sterck, E.H.M. & Goossens, B.M.A. The meaning of “macaque” facial expressions. *Proc. Natl. Acad. Sci. U.S.A.* **105**, E71-E71 (2008). doi: 10.1073/pnas.0806462105

- Striemer, C.L., Whitwell, R.L. & Goodale, M.A. Affective blindness in the absence of input from face processing regions in occipital-temporal cortex. *Neuropsychol.* **128**, 50-57 (2019).
- Sugase, Y., Yamane, S., Ueno, S. & Kawano, K. Global and fine information coded by single neurons in the temporal visual cortex. *Nature.* **400**, 869–873 (1999).
- Tamietto, H., Castelli, L., Vighetti, S., Perozzo, P., Geminiani, G., Weiskrantz, L. & de Gelder, B. Unseen facial and bodily expressions trigger fast emotional reactions. *Proc. Natl. Acad. Sci. USA* **106**, 17661-17666 (2009).
- Tamietto, H. & de Gelder, B. Neural bases of the non-conscious perception of emotional signals. *Nat. Rev. Neurosci.* **11**, 697-709 (2010).
- Taubert, J., Wardle, S.G. and Ungerleider, L.G. What does a “face cell” want?’. *Progress in neurobiology*, **195**, p.101880 (2020).
- Tokui, S., Oono, K., Hido, S. & Clayton, J. Chainer: a next-generation open source framework for deep learning. *Proc. Workshop on machine learning systems (LearningSys) in 29th annual conference on neural information processing systems.* **5**, 1-6 (2015).
http://learningsys.org/papers/LearningSys_2015_paper_33.pdf.
<https://github.com/chainer/chainer/releases/tag/v3.0.0> (version, 3.0.0; release, Oct 17, 2017)
- Tsao, D.Y. & Livingstone, M.S Mechanisms of face perception. *Annu. Rev. Neurosci.* **31**, 411-437 (2008).
- Ueda, Y., Nunoi, M., & Yoshikawa, S. Development and validation of the Kokoro Research Center (KRC) facial expression database. *Psychologia* **61**, 221-240 (2019).
- Ungerleider, L.G. & Mishkin, M. Two cortical visual systems. In *Analysis of Visual Behavior*. Edited by Ingle, D.G., Goodale, M.A. & Mansfield, R.J.Q. MIT Press: Cambridge, pp. 549-586 (1982)

- Uptdike, B.V. Characteristics of unit responses in superior colliculus of the Cebus monkey. *J. Neurophysiol.* **37**, 896-909 (1974).
- Van den Bergh, G., Zhang, B., Arckens, L. and Chino, Y.M. Receptive-field properties of V1 and V2 neurons in mice and macaque monkeys. *J. Comp. Neurol.* **518**, 2051–2070 (2010).
- Vaziri, S., Calson, E.T., Wang, Z. & Connor, C.E. A channel for 3D environmental shape in anterior inferotemporal cortex. *Neuron* **84**, 55-62 (2014).
- Verhoef, B-E., Vogels, R. & Janssen, P. Binocular depth processing in the ventral visual pathway. *Phil. Trans. R. Soc. B* **371**, 20150259 (2016).
- Vuilleumier, P., Armony, J.L., Driver, J. & Dolan, R.J. Distinct spatial frequency sensitivities for processing faces and emotional expressions. *Nat. Neurosci.* **6**, 624–631 (2003).
- Wolfram Research Inc., Wolfram Alpha Computational intelligence AI, available from<<http://wolfram Alpha.com>>, accessed on 19 May, 2023.
- Wallace, M. T., McHaffie, J. G. & Stein, B. E. Visual response properties and visuotopic representation in the newborn monkey superior colliculus. *J. Neurophysiol.* **78**, 2732–2741 (1997).
- Waller, B.M., Julle-Daniere, E. & Micheletta, J. Measuring the evolution of facial ‘expression’ using multi-species FACS. *Neurosci. Biobehav. Rev.* **113**, 1-11 (2020). doi: 10.1016/j.neubiorev.2020.02.031
- Whalen, P.J., Rauch, S.L., Etcoff, N.L., McInerney, S.C., Lee, M.B. & Jenike, M.A. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *J. Neurosci.* **18**, 411-418 (1998).
- Yamins, D.L.K. & DiCarlo, J.J. Using goal-driven learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356-365 (2016).

Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D. & DiCarlo, J.J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619-8624 (2014).

Personal Information

Name: Chanseok Lim
2018 – 2023, Ph D course graduate student

Affiliations: 2022 – 2023, Perceptual and Cognitive Neuroscience Laboratory, Graduate School of Frontier Biosciences, Osaka University
2018 – 2022, Laboratory for Cognitive Neuroscience, Graduate School of Frontier Biosciences, Osaka University

Address: 1-3 Machikaneyama, Toyonaka, Osaka, 560-8531, Japan

Thesis title: Convolutional Neural Network Models for Subcortical Processing of Facial Expression

Supervisors: Profs. Shinji Nishimoto and Ichiro Fujita

Committee: Profs. Shinji Nishimoto (chair), Shigeru Kitazawa, Takeshi Yagi, and Assoc. Prof. Hiroshi Tamura

Education: 2014 – 2018, School of Engineering Science, Osaka University

Bibliography

Publication (international)

1. Lim, C., Inagaki, M., Shinozaki, T. and Fujita, I. Analysis of convolutional neural networks reveals the computational properties essential for subcortical processing of facial expression. *Scientific Reports*, 13(1), 10908 (2023).

Publication (domestic)

1. 林燦碩, 稲垣未来男, 篠崎隆志, 藤田一郎. 畳み込みニューラルネットワークを用いた皮質下経路の粗い情報処理の説明. 信学技報, vol. 121, no. 271, NC2021-28, pp. 1-6, 2021 年 11 月.
2. 小松優介, 稲垣未来男, 林燦碩, 篠崎隆志, 藤田一郎. 顔表情弁別を行う畳み込みニューラルネットワークの内部における空間周波数特性. 信学技報, vol. 118, no. 367, NC2018-29, pp. 5-10, 2018 年 12 月.
3. 林燦碩, 稲垣未来男, 小松優介, 篠崎隆志, 藤田一郎. CNN における扁桃体細胞類似特性獲得のための視覚体験的学習法. 信学技報, vol. 118, no. 322, NC2018-24, pp. 5-10, 2018 年 11 月.

Presentation (international)

1. Lim, C., Inagaki, M., Shinozaki, T. and Fujita, I. Why is Visual Information Conveyed by the Subcortical Pathway Coarse? The 7th CiNet Conference: New horizons in brain mapping, online, 2022/02/01.

Presentation (domestic)

1. 稲垣未来男, 佐藤優, 渡辺直樹, 林燦碩, 小松優介, 篠崎隆志, 藤田一郎. 大脳皮質経路と皮質下経路が表情認識に与える影響：畳み込みニューラルネットワークと心理実験による検討. 日本視覚学会 2022 年夏季大会, 金沢大学, 2022 年 9 月 5 日.
2. 林燦碩, 稲垣未来男, 篠崎隆志, 藤田一郎. 畳み込みニューラルネットワークを用いた皮質下経路の粗い情報処理の説明. ニューロコンピューティング研究会, オンライン開催, 2021 年 11 月 26 日.
3. 赤星宏知, 稲垣未来男, 林燦碩, 小松優介, 篠崎隆志, 藤田一郎. 大脳皮質経路と皮質下経路の表情表現：畳み込みニューラルネットワークを用いた比較. 日本視覚学会 2021 年夏季大会, オンライン開催, 2021 年 9 月 22 日.

4. 渡辺直樹, 稲垣未来男, 林燦碩, 小松優介, 篠崎隆志, 藤田一郎. 大脳皮質と皮質下における恐怖表情処理の比較: 畳み込みニューラルネットワークによる検討. 日本視覚学会 2020 年夏季大会, オンライン開催, 2020 年 9 月 17 日.
5. 小松優介, 稲垣未来男, 林燦碩, 篠崎隆志, 藤田一郎. 顔表情弁別を行う畳み込みニューラルネットワークの内部における空間周波数特性. ニューロコンピューティング研究会, 名古屋工業大学, 2018 年 12 月 15 日.
6. 林燦碩, 稲垣未来男, 小松優介, 篠崎隆志, 藤田一郎. CNN における扁桃体細胞類似特性獲得のための視覚体験的学習法. ニューロコンピューティング研究会, 京都大学, 2018 年 11 月 22 日.
7. 稲垣未来男, 林燦碩, 小松優介, 篠崎隆志, 藤田一郎. 大脳皮質経路と皮質下経路における顔処理のモデル化の試み. 日本視覚学会 2018 年夏季大会, 文部科学省研究交流センター, 2018 年 8 月 2 日.
8. Lim, C., Inagaki, M., Komatsu, Y., Shinozaki, T., and Fujita, I. A CNN model for the subcortical pathway acquires amygdala neuron-like properties when trained in human-like experiences. 第 41 回日本神経科学大会, 神戸コンベンションセンター, 2018 年 7 月 28.
9. Komatsu, Y, Inagaki, M., Lim, C., Shinozaki, T., and Fujita, I. Internal representation of convolutional neural network in classifying facial expression. 第 41 回日本神経科学大会, 神戸コンベンションセンター, 2018 年 7 月 28 日.

Grants (completed)

Apr. 2020 – Mar. 2023 Grant-in-Aid for JSPS Fellows (DC1). Funding Agency: Japan Society for the Promotion of Science (JSPS). Project number: 20J22553. Title: 霊長類扁桃体への迅速な脅威信号伝達の視覚経路における顔表情検出原理の解明. Role: PI. Amount: 200,000 JPY/month + 3,100,000 JPY (Direct: 3,100,000 JPY).

Awards

1. 林 燦碩, 稲垣 未来男, 篠崎 隆志, 藤田 一郎. 優秀研究賞. 日本神経回路学会. 2022 (2021 年度).
2. Chanseok Lim. Young Researcher Award (IEICE Neurocomputing). IEEE Computational Intelligence Society Japan Chapter. 2022 (2021 年度).
3. 小松 優介, 稲垣 未来男, 林 燦碩, 篠崎 隆志, 藤田 一郎. 最優秀研究賞. 日本神経回路学会. 2019 (2018 年度).
4. 林 燦碩. 大阪大学楠本賞 (学科首席). 大阪大学. 2018 (2017 年度).