Mapping between the Human Visual System and Two-stream DCNNs in Action Representation

Yujia Peng[1,2,6,7*] yujia_peng@pku.edu.cn
Xizi Gong[1*] gongxizi0730@pku.edu.cn
Hongjing Lu[7,8] hongjing@ucla.edu
Fang Fang[1,3,4,5] ffang@pku.edu.cn

[1] School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health
[2] Institute for Artificial Intelligence [3] Key Laboratory of Machine Perception, Ministry of Education,
[4] IDG/McGovern Institute for Brain Research [5] Peking-Tsinghua Center for Life Sciences
Peking University, Beijing, China
[6] National Key Laboratory of General Artificial Intelligence, BIGAI, Beijing, China
[7] Department of Psychology [8] Department of Statistics, University of California, Los Angeles, USA
* Equal contribution

Abstract

Deep convolutional neural networks (DCNNs) have been found to demonstrate hierarchical mapping to human brain regions on tasks such as object recognition. However, it remains unclear if such hierarchical mapping also applies to action recognition, which involves dynamic visual information processing. Here, we compared action representations of two-stream DCNNs to the human visual system. Five visual areas that are associated with object and action processing were selected. Nine human action categories were adopted from three semantic classes to examine the action representations of both DCNNs and human visual areas. In two fMRI experiments, actions were presented in the forms of computer-rendered videos and point-light biological motion videos. Results showed that although two-stream DCNNs demonstrated hierarchical representations of actions as layers grow deeper, DCNNs lack a hierarchical mapping to human visual areas. Consistently across different video displays and DCNN pathways, only the top DCNN layers demonstrated high similarity to representations in the human visual system. The results suggest that the dynamic representations of human actions may be different in DCNNs compared to the human visual system, even after big-data training.

Keywords: action perception; deep convolutional neural networks; biological motion; fMRI; hierarchical representation

Introduction

One of the most robust and sophisticated abilities supported by the human visual system is the recognition of human actions. In daily life, humans can readily recognize actions despite changes in body forms and appearance and with various types of visual noise. Even in highly impoverished and rarely observed stimuli such as point-light videos (Johansson, 1973), the human visual system can still recognize actions despite visual noise (Lu, 2010; Neri et al., 1998) and efficiently identify attributes of an actor (for example, Peng et al., 2017; Peng et al., 2021; Pollick et al., 2002; Thurman & Lu, 2016; Thurman & Lu, 2014). It is essential to understand how the human brain achieves sophisticated semantic-level representations of human actions.

Over several decades, psychophysical and neuroscience research has provided evidence suggesting that recognition of biological motion may be supported by both the spatial structure of body forms and motion information (Beintema & Lappe, 2002; Cutting et al., 1988; Lange et al., 2006; Pinto & Shiffrar, 1999; Theusner et al., 2011; van Boxtel & Lu, 2015). In particular, fMRI experiments have shown that point-light videos activate not only motion-selective regions such as MT/MST, but also a projection from the primary visual cortex to the inferotemporal cortex that processes object appearance information (Grossman & Blake, 2002). In addition, the extrastriate body area (EBA) has been implicated in recognizing human body forms (Downing, 2001; Lingnau & Downing, 2015). Finally, numerous studies have established that the posterior superior temporal sulcus (pSTS) is a region supporting biological motion perception, integrating motion processing and appearance processing (Grossman et al., 2005, 2010; Grossman & Blake, 2001, 2002; Thurman et al., 2016; Vaina et al., 2001).

Inspired by the aforementioned findings on point-light videos and the hierarchical processing of static information as in object recognition, Giese and Poggio (2003) developed a parsimonious model for action recognition with two parallel processing streams: a "what" pathway and a "where" pathway. The "what" pathway is specialized for analyzing body forms in static image frames. The "where" pathway is specialized for processing optic flow or motion information. Both pathways comprise a hierarchy of feature detectors with increasing receptive fields and complexities in encoding form or motion patterns. Building upon previous works, Simonyan and Zisserman (2014) developed two-stream deep convolutional neural networks (DCNNs) (Krizhevsky et al., 2012; Lecun et al., 1998) for action recognition. The two-stream DCNN consists of two DCNNs: a spatial DCNN that processes appearance information taking pixel-level intensity as the input, and a temporal DCNN that processes motion information taking optical flow as the input. The two-stream DCNN performed well on action classification for two challenging datasets: UCF-101 (Soomro et al., 2012) and HMDB-51 (Kuehne et al., 2011). There are other deep learning models developed for video recognition. For example, slowfast networks by Feichtenhofer and colleagues (2019) also use two pathways for action recognition from videos, but both pathways operate on a clip of video as a spatiotemporal volume with different frame rates. The architecture in slowfast networks does not clearly map to the

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

"what" and "where" pathways in the brain. The other popular network is a two-stream inflated 3D convnet (I3D) developed by Carreira & Zisserman (2017). The I3D model is built on the basis of the inception-V1 network structure and includes 9 inception layers. In contrast with many studies on comparing human visual regions with the DCNN models (such as AlexNet), there exists little evidence on the correspondence between visual areas and inception layers. Hence, this paper focuses on the two-stream DCNNs as an extension of standard DCNN models (such as AlexNet), which have rich literature on human and model comparisons.

Despite tremendous recent advances in AI, human intelligence is still far more adept at understanding the observed dynamic information in the real world. It remains unclear whether deep neural networks contain similar representations as human brains. In object recognition, numerous neural imaging studies have reported a hierarchical DCNN-brain correspondence: the representation of DCNN layers can predict image-driven responses along the ventral visual stream and reveals representations of increasingly complex information as the layers go deeper (Cadena et al., 2019; Cadieu et al., 2014, 2014; Cichy et al., 2016, 2017; Eickenberg et al., 2017; Güçlü & van Gerven, 2015; Hong et al., 2016; Khaligh-Razavi et al., 2017; Khaligh-Razavi & Kriegeskorte, 2014; Seeliger et al., 2018; Yamins et al., 2013; Yamins et al., 2014). However, few studies have focused on dynamic stimuli such as motion and action stimuli. Comparisons between the two-stream DCNN and the human brain may open a window to reveal how action information gradually unfolds across regions of interest (ROIs).

In the current study, we examined the mapping between representations of two-stream DCNNs and human visual areas on human action perception. We selected five ROIs along the two-stream visual pathways: the primary visual cortex (V1) for low-level visual information processing, middle temporal/medial superior temporal (MT+) for motion processing, lateral occipital complex (LOC) for object perception, extrastriate body area (EBA) for human body processing, and posterior superior temporal sulcus (pSTS) known for biological motion perception and theory-of-mind. We use both computer-rendered videos and decontextualized point-light videos to examine the processing of human actions in different presentation formats and generalization ability. Point-light videos remove detailed body-shape and contextual information and only keep the motion trajectories of major joints in actions, whereas computer-rendered videos present the same actions with greater ecological validity, rendering the actions with human avatars.

If the process of action recognition resembles the process of object recognition, we would expect to see a hierarchical mapping between the representation of DCNN layers and human brain regions, such that early layers of DCNNs demonstrate representations more similar to V1 and MT+, and later DCNN layers demonstrate increasingly similar representations to layers such as pSTS. Additionally, we expected to find correspondences between the spatial DCNN and the "what" visual pathway for form processing, and between the temporal DCNN and the "where" visual pathway for motion processing.

Model Structure and Training

To investigate action representations in DCNN, we selected a two-stream DCNN model (Figure 1) with an architecture based on neurophysiological and computational studies in the biological motion literature (Giese & Poggio, 2003). Specifically, biological motion processing involves both form and motion pathways and integrates the two types of information at action-sensitive regions, presumably in the temporal lobe. The two-stream DCNN takes the same two types of information as inputs to classify a video into action categories. One source of information is the pixel-level appearance of moving bodies in a sequence of static images, which provide inputs to a spatial DCNN. The other source of information is motion represented by optical flow fields (Horn & Schunck, 1981), which provide inputs to a temporal DCNN. Both the spatial and temporal DCNNs contain 5 convolutional layers followed by 3 fully-connected layers. At the 5 convolutional layers, a two-stream DCNN model combines the spatial and motion processes to achieve a fusion of decisions.

The DCNN models were trained to perform an action classification task with the 15 categories using naturalistic videos in the Human 3.6M dataset (Ionescu et al., 2014). We followed a two-phase protocol to train the network as developed by Feichtenhofer, Pinz, and Zisserman (2016). We first trained the single-stream networks (i.e., the spatial DCNN and the temporal DCNN) independently with the task of 15-category action recognition. Then activities from the conv5 layers of these two trained single-stream DCNNs were concatenated as inputs to train the fusion network in the two-stream DCNN. These 15 categories include giving directions, discussing something with someone, eating, greeting someone, phoning, posing, purchasing (i.e., hauling up), sitting, sitting down, smoking, taking photos, waiting, walking, walking a dog, and walking together.

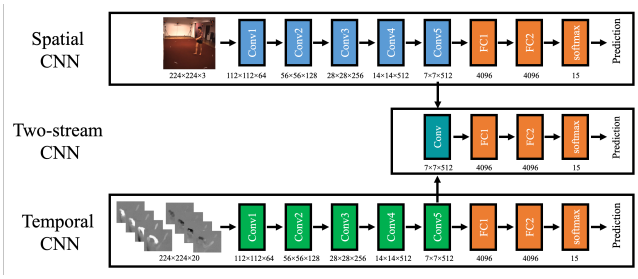

Figure 1: The architectures of the spatial DCNN, the temporal DCNN, and the two-stream DCNN.

Methods

Participants

Twelve subjects (7 female, age (M±SD) = 21.17±1.85) participated in the study, with half presented with Computer-

rendered videos and the other half presented with biological motion stimuli. Each subject went through five fMRI sessions on separate days to maximize the robustness of acquired neural representations. All subjects were right-handed and had normal or corrected-to-normal vision. They had no known neurological or visual disorders and gave written, informed consent in accordance with the procedures and protocols approved by the human subject review committee.

Stimuli

Action stimuli were generated from the Carnegie Mellon University Motion Capture Database. We adopted the same superordinate categories from a study by Dittrich (1993): human actions can be considered as falling into three semantic classes: locomotory, instrumental, and social actions. We selected 3 actions for each class (Locomotory action: jumping, running, & walking; Instrumental action: ball bouncing, playing an instrument, & golf swing; Social action: dancing, greeting, & showing directions; Figure 2A), each with 4 different motion-tracking instances, resulting in 9 cation categories and 36 action instances in total. Point-light videos (Figure 2B) were processed using the Biological Motion Toolbox (van Boxtel & Lu, 2013). Autodesk Maya® was used to render motion-tracking data with human avatars to generate computer-rendered videos. Hence, body movements in computer-rendered and point-light videos were the same, despite the differences in body shape and visual appearance in these two displays.

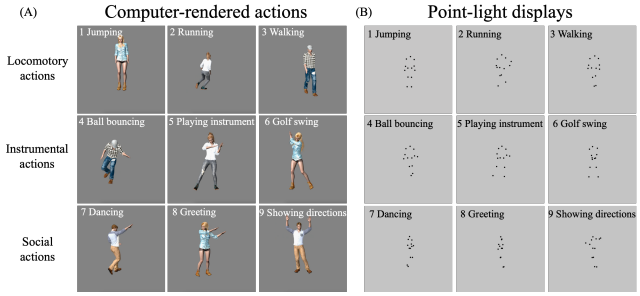

Figure 2: Sample frames of the nine action categories selected from the CMU motion capture database, falling into three semantic classes of locomotory, instrumental, and social actions. Actions were presented in (A) Computer-rendered actions or (B) point-light biological motion displays.

Procedure

The whole experiment was conducted across five days. As shown in Figure 3A, on day 1, subjects completed the behavioral practice, the structural scan, and the localizer tasks. Subjects first went through a behavioral practice session, during which they were trained to classify actions into the three semantic classes. Subjects went through two practice runs, where all 36 videos were presented during each run. Overall, the subjects all reached near-perfect behavioral accuracy. After the behavioral practice, subjects underwent an MRI session including structural scans and the localizer tasks, aiming to define five ROIs, namely V1 (Engel et al., 1997), MT (Watson et al., 1993), LOC (Malach R et al., 1995), EBA (Downing et al., 2001), and pSTS (Grossman et al., 2000).

In the following four days, after finishing a structural scan, subjects performed eight runs of the action classification task each day, resulting in 32 runs. Each run started with 10s of fixation, followed by 36 trials of action presentations, each presented for 3 seconds, interleaved by a period for response and jitter of 3, 5, or 7s, ending with another 10s of fixation. During the response period, subjects were asked to judge the semantic class of the action by pressing one of the three buttons on the response box.

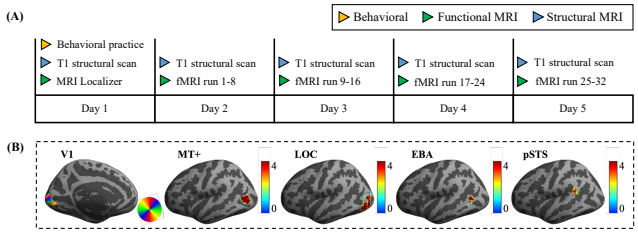

Figure 3: (A) Experimental procedure across five days. (B) Illustrations of ROI positions from one subject.

Data analysis

Representational similarity analysis (RSA) was used to compare neural representations with the DCNN representations. Specifically, for each layer of the DCNN, we extracted condition-specific neuron activation values of video clips (every 10 frames) in each action instance. For convolutional layers, we used a max-pooling approach to take the maximum response value from each 2D response field. First, features for video clips in each action instance were concatenated into a vector. Then, for each pair of actions, we computed the Euclidean distances dissimilarity between the model activation pattern vectors, yielding a 36 × 36 DCNN representational dissimilarity matrix (RDM), summarizing the representational dissimilarities for each model layer.

Similarly, a correlation-based approach was used to compute RDMs for ROIs. For each ROI, condition-specific beta-value activation patterns of voxels were concatenated into vectors. We then calculated the correlation-based dissimilarity between t-value patterns for every pair of conditions within the ROI, leading to a 36 × 36 ROI RDM indexed in rows and columns by the compared actions.

We compared layer-specific model representations to region-specific brain representations by calculating Spearman's correlations between the lower half of the DCNN and ROI RDMs, excluding the diagonal. The comparison was done on a single-subject basis. We estimated the noise ceiling for each ROI, and the DCNN-ROI correlations were normalized by dividing the raw correlation coefficients with the corresponding noise ceiling. The noise ceiling was defined as $\frac{1}{N}\sum_{i=1}^{N} r(v_i, \bar{v})$, where v_i represents each subject's RDM, $\bar{v}$ represents the averaged RDM across subjects, and r stands for Spearman's correlation coefficient (Neli et al., 2014; Khaligh-Razavi et al., 2018).

To provide theoretical guidance to the representation of action categories, we made a full-knowledge design matrix of RDMs, as shown in Fig. 4E. The full-knowledge design matrix assumes higher similarities between videos within one main semantic class category on top of similarities within one action category (e.g., jumping, walking, and running are all locomotory actions). DCNN and ROI RDMs were compared to design matrices to investigate whether model layers and ROIs demonstrate low-level representations based on action-specific visual features or semantic-level action representations beyond visual similarities of actions from the same action category.

Results

DCNN Action Representations

We first examined the representations in DCNN layers and how well they discriminate action categories and semantic classes. As shown in Fig. 4A and 4B for both computer-rendered videos and point-light videos, all three DCNN pathways demonstrated more information about individual action categories with the increase of layers, revealed by more apparent mini-blocks pattern along the diagonals. Spatial DCNN processing appearance information demonstrates rather different dissimilarity patterns for computer-rendered videos from dissimilarity patterns from point-light videos. This difference in the spatial DCNN was likely due to distinct appearance differences between the two displays. In contrast, the temporal DCNN processing optical-flow information was less impacted by different display formats, showing similar diagonal block patterns organized by action categories.

Correlations between DCNN layer RDMs and the design matrix are shown in Fig. 5C and 5D. Regression analyses revealed significant linear relationships between layers and correlation coefficients of the temporal DCNN and the full-knowledge design matrix (Computer-rendered: b=0.060, $t(4)$=6.15, p=.004; Point-light: b=0.058, $t(4)$=3.86, p=.018). The results further confirmed the pattern observed in the decoding analysis: as DCNN layers go deeper, the action representation of corresponding layers increasingly resembles the full-knowledge design matrix. We also found that the regression between spatial DCNN layers and design matrices was only significant for computer-rendered videos (b=0.024, $t(4)$=3.59, p=.023) and not significant for point-light videos, suggesting body form cues can still contribute to action recognition given the similarity between computer avatars and humans in naturalistic videos.

Action Representations in Human Brains

We next examined the RDMs of selected ROIs along the visual pathways. As shown in Fig. 4C and 4D, all five ROIs demonstrated diagonal block patterns indicating discrimination of action categories and semantic classes.

To investigate how well the representations discriminate action categories and semantic classes, correlations between ROI RDMs and the full-knowledge design matrix were calculated (Fig. 5C and 5D). For correlations to the full-knowledge design matrix, paired-sample t-tests showed significant contrasts between ROIs. For computer-rendered videos, V1 yielded significantly lower correlations than MT+, LOC, and EBA. Additionally, MT+, LOC, and EBA produced significantly greater correlations than pSTS, surviving Bonferroni corrections (p corrected < 0.05). For point-light videos, V1 produced significantly lower correlations to the full-knowledge design matrix than EBA and survived the Bonferroni correction (p corrected < 0.05).

ROI-DCNN RDM correlations

The normalized correlations between DCNN and ROI action representations were shown in Fig. 6. Most correlations between ROI and DCNN RDMs were significant, except for early spatial DCNN layers (e.g., convolutional layers 1 to 3). Furthermore, for all ROIs, as the DCNNs go to deeper layers, the correlations between RDMs increased for both spatial and temporal DCNNs and the two-stream DCNN.

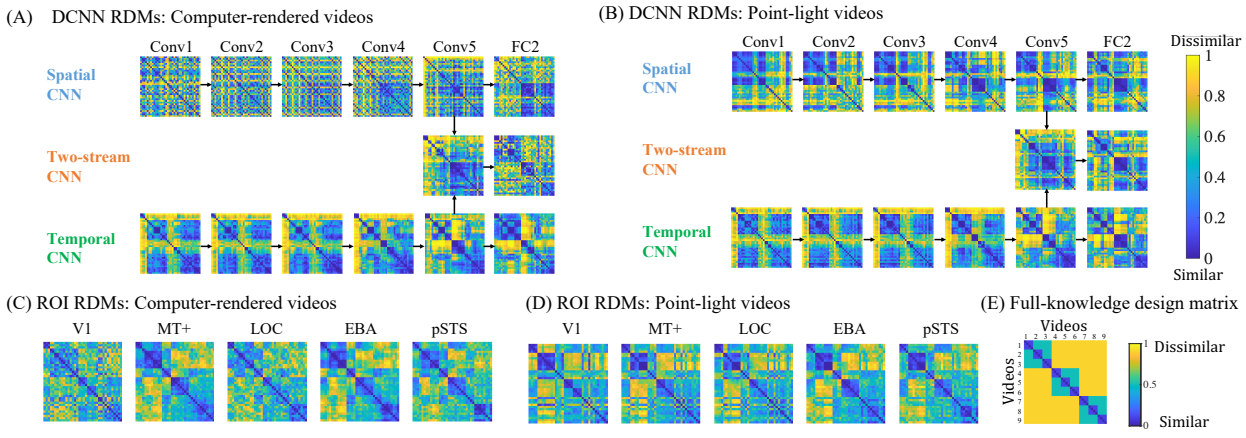

Figure 4: DCNN RDMs with (A) Computer-rendered videos and (B) point-light videos, and averaged ROI RDMs with (C) Computer-rendered videos and (D) point-light videos across subjects. (E) The full-knowledge design matrix RDM. The action categories were labeled as 1-9, corresponding to label numbers in Fig. 2.

A repeated-measures ANOVA was conducted with ROIs, DCNN networks (spatial and temporal DCNNs), and layers as within-subject variables. Results showed significant main effects of DCNN in both experiments (*p*s<.001), suggesting that the temporal DCNN yielded greater correlations to ROI RDMs in general compared to the spatial DCNN. Results also showed significant main effects of layers (*p*s<.05) in both experiments. Thus, in contrast to the expected pattern that different ROIs would reach the maximum correlation with different DCNN layers, all ROIs yielded the greatest representational similarity to the later layers of the DCNN, namely the Conv5 layer and the fully-connected FC2 layer.

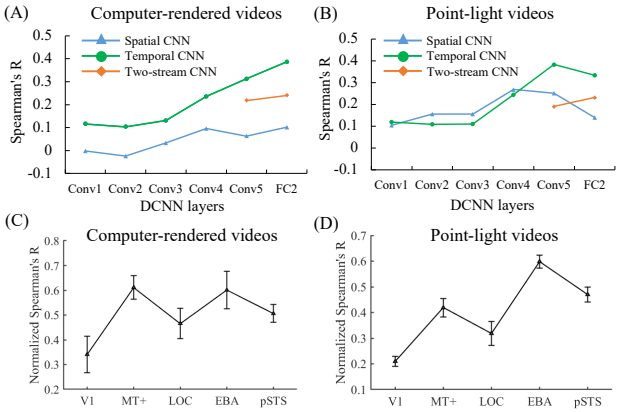

Figure 5: Correlations between DCNN and the full-knowledge design matrix of Computer-rendered videos (A) and point-light videos (B). Normalized correlations between ROIs in the human visual system and the full-knowledge design matrix of computer-rendered videos (C) and point-light videos (D).

Searchlight analyses

To identify brain areas with action representations similar to those of the DCNN layers, we used a spatially unbiased volume-based searchlight approach. For each subject, we constructed fMRI RDMs for each voxel (3-voxel radius) based on the local activity patterns. We then correlated each voxel's RDM with the layer-specific DCNN RDMs, generating a continuous spatial map of similarity for each DCNN layer. The searchlight approach also revealed a gradually increasing correspondence between the DCNN layers and human cortices as layers go deeper. The strongest correlations were observed between the Conv5 layer and FC2 layers of DCNNs and brain regions such as anterior intraparietal sulcus (IPS) and superior parietal lobule (SPL). These results further suggested that the anterior parietal cortex may play a crucial role in action processing.

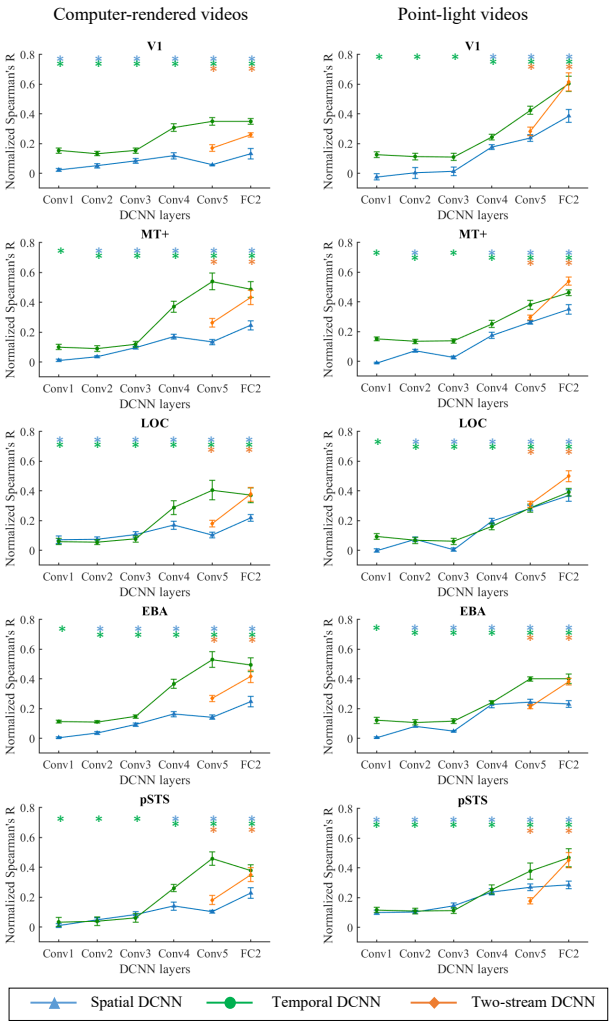

Figure 6: Normalized correlations between ROI RDMs and DCNN RDMs for computer-rendered videos and for point-light videos. Blue, green, and orange lines correspond to the spatial, temporal, and two-stream DCNNs. Error bars indicate standard deviations across subjects. The stars above bars indicate the significance of ROI-DCNN correlations of layers with corresponding colors (*p*s < 0.05, FDR-corrected).

Discussion

DCNN Action Representations

In the present study, we examined the representation mapping between DCNN layers and visual regions in human brains supporting action perception. For DCNN, results showed a robust hierarchical representation of human actions where representational dissimilarity matrices yielded clearer clustering patterns that discriminate action categories as layers go deeper. For human visual areas, all selected regions were able to discriminate action categories. However, comparisons between DCNN layers and ROIs across two experiments consistently revealed a lack of hierarchical correspondences, as DCNN layers yielded the highest similarity to brain representations at later layers

(convolutional layer 5 or fully connected layer 2), regardless of the type of action displays, the DCNN pathways, or ROIs.

The finding of hierarchical representations of two-stream DCNNs was consistent with previous evidence that after big-data training, early layers of DCNN capture low-level visual features that resemble receptive fields of early visual areas, and later DCNN layers capture informative visual patterns with increasing visual complexity for visual recognition (e.g., Yamins et al., 2014). The increasingly categorical representation was expected in deep neural networks. As proposed by Saxe et al. (2019), it is possible that the ability of DCNNs to capture abstract semantic knowledge consisting of useful categories may be inherent in the deep connection structure. But it is also possible that abstract categorical information is highly associated with some visual features that the networks learn to capture in feature extraction.

However, the lack of a hierarchical mapping between DCNN layers and brain regions was at odds with the findings in object recognition with static images. One possibility may be due to differences in the nature of the tasks. Previous research focused mostly on object recognition, and the current study targeted dynamic visual stimuli of human actions. Object recognition can possibly be achieved in a single stream of feedforward processing where low-level visual features are extracted and integrated to capture the essence of complex object patterns in later layers. This feedforward process emulates bottom-up visual processing in the human brain. However, the processing of dynamic visual information may require large spatial-temporal windows to accumulate information over time and space. Action recognition unfolds over time during which communications between brain regions happen, and iterations of bottom-up feature extraction and top-down influences both may play important roles in making decisions and responses (Lu, Tjan & Liu, 2006). Even though the two-stream DCNN provided a qualitative account of some behavioral results observed in human action perception, DCNNs are limited to only operating in a purely bottom-up manner and lack top-down regulation apparent in human brains (Peng, Lee, et al., 2021).

The current results cannot rule out hierarchical structures of action representations in the human brain, but may support a fast unfolding of action perception over time. Neuroimaging techniques with a greater temporal resolution have provided evidence that supports the fast unfolding of action representations in human brains. Previous magnetoencephalography (MEG) studies showed that the recognition of human social interactions may involve different visual mechanisms than simple feedforward pattern recognition (Isik et al., 2020). Different types of human social interactions can be decoded at around 500 ms after the onset of videos, which is substantially later than the visual processing of objects, faces, emotions, gestures, and actions. For example, object pattern recognition can be decoded within 100 ms of the image onset (e.g., Carlson et al., 2013; Isik et al., 2014). Face perception elicits the signature N170 response at around 170 ms after face image onset (Bentin et al., 1996), while many facial properties such as age, gender, and identity can be decoded even earlier (Dobs et al., 2019). Communicative gestures (Redcay & Carlson, 2015) and single-person actions can be decoded as early as 200 ms (Isik et al., 2018). Thus, unlike visual processing of static images or single-agent movements, inference of intentions from human social interactions may involve the recognition of high-level semantics and relational reasoning that go beyond visual pattern recognition.

A few limitations can be addressed in future studies to further illuminate action representations underlying human and artificial neural networks. First, while fMRI provided a good spatial resolution to reveal specificities of ROI representations along the visual pathway, it lacks the temporal resolution to reveal the neural dynamics across time. Future studies can use MEG or EEG to investigate how representations of human actions unfold over time, and whether DCNN representations demonstrate hierarchical relationships to evolving brain representations. Secondly, the current fMRI paradigm was based on a classification task that may not require social cognitive processes, such as theory-of-mind, which are involved in daily action processing. In addition, we only examined a small number of categories of actions; future research can expand to a larger variety of action stimuli and semantic classes. Lastly, the correspondence between DCNNs and human neural dynamics on a finer scale can be investigated. The current study was limited to several classic visual regions, but future studies can explore whole-brain neural responses and connectivities between brain regions.

In summary, the current study adopted the two-stream DCNN trained with big data to examine the relationship between action representations of artificial neural networks and human visual pathways. The findings indicate a lack of hierarchical relationship between DCNN layers and human visual regions. Instead, while the DCNN layers demonstrate increasingly high-level representations, they may not resemble the efficient representations in the human brain. The current study provides evidence that deep neural networks open a window for understanding the dynamic visual processes in human brains. Human neuroimaging studies can also reveal limitations and provide guidance for developments in artificial intelligence.

Acknowledgements

This work was funded by National Science and Technology Innovation 2030 Major Program (2022ZD0204802, 2022ZD0204804), the NSFC (31930053), and the Beijing Academy of Artificial Intelligence to F.F. We thank Qi Xie for helping on generating the computer-rendered action stimuli, Tianmin Shu for assistance on DCNN model training, Junshi Lu for providing helpful advice on ROI selection and data analysis, and all the participants for their contribution to this research.

References

Beintema, J. A., & Lappe, M. (2002). Perception of biological motion without local image motion.

Proceedings of the National Academy of Sciences of the United States of America, *99*(8), 5661–5663.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, *8*(6), 551–565.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Computational Biology*, *15*(4), e1006897.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Computational Biology*, *10*(12), e1003963.

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1–1.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299-6308).

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, *153*, 346–358.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.

Cutting, J. E., Moore, C., & Morrison, R. (1988). Masking the motions of human gait. *Perception & Psychophysics*, *44*(4), 339–347.

Dittrich, W. H. (1993). Action Categories and the Perception of Biological Motion. *Perception*, *22*(1), 15–22.

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications*, *10*(1), 1–10.

Downing, P. E. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cerebral cortex (New York, NY: 1991)*, *7*(2), 181-192.

Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In P*roceedings of the IEEE/CVF international conference on computer vision* (pp. 6202-6211).

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience*, *4*(3), 179–192.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., & Blake, R. (2000). Brain Areas Involved in Perception of Biological Motion. *Journal of Cognitive Neuroscience*, *12*(5), 711–720.

Grossman, E. D., Battelli, L., & Pascual-Leone, A. (2005). Repetitive TMS over posterior STS disrupts perception of biological motion. *Vision Research*, *45*(22), 2847–2853.

Grossman, E. D., & Blake, R. (2001). Brain activity evoked by inverted and imagined biological motion. *Vision Research*, *41*(10), 1475–1482.

Grossman, E. D., & Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, *35*(6), 1167–1175.

Grossman, E. D., Jardine, N. L., & Pyles, J. A. (2010). FMR-adaptation reveals invariant coding of biological motion on human STS. *Frontiers in Human Neuroscience*, *4*.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622.

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*(1–3), 185–203.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1325–1339.

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102.

Isik, L., Mynick, A., Pantazis, D., & Kanwisher, N. (2020). The speed of human social interaction perception. *NeuroImage*, *215*, 116844.

Isik, L., Tacchetti, A., & Poggio, T. (2018). A fast, invariant representation for human action in the visual system. *Journal of Neurophysiology*, *119*(2), 631–640.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211.

Khaligh-Razavi, S.-M., Cichy, R. M., Pantazis, D., & Oliva, A. (2018). Tracking the Spatiotemporal Neural Dynamics of Real-world Object Size and Animacy in the Human Brain. *Journal of Cognitive Neuroscience*, *30*(11), 1559–1576.

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., & Kriegeskorte, N. (2017). Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, *76*, 184–197.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain

IT Cortical Representation. *PLOS Computational Biology*, *10*(11), e1003915.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: a large video database for human motion recognition. *2011 International Conference on Computer Vision*, 2556–2563.

Lange, J., Georg, K., & Lappe, M. (2006). Visual perception of biological motion by form: A template-matching analysis. *Journal of Vision*, *6*(8), 6.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lingnau, A., & Downing, P. E. (2015). The lateral occipitotemporal cortex in action. *Trends in Cognitive Sciences*, *19*(5), 268–277.

Lu, H. (2010). Structural processing in biological motion perception. *Journal of Vision*, *10*(12), 13–13.

Lu, H., Tjan, B. S., & Liu, Z. (2006). Shape recognition alters sensitivity in stereoscopic depth discrimination. *Journal of Vision*, *6*(1), 7-7.

Malach. R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135–8139.

Neri, P., Morrone, M. C., & Burr, D. C. (1998). Seeing biological motion. *Nature*, *395*(6705), 894–896.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*(4), e1003553.

Peng, Y., Lee, H., Shu, T., & Lu, H. (2021). Exploring biological motion perception in two-stream convolutional neural networks. *Vision Research*, *178*, 28–40.

Peng, Y., Lu, H., & Johnson, S. P. (2021). Infant perception of causal motion produced by humans and inanimate objects. *Infant Behavior and Development*, *64*, 101615.

Peng, Y., Thurman, S., & Lu, H. (2017). Causal Action: A Fundamental Constraint on Perception and Inference About Body Movements. *Psychological Science*, *28*(6), 798–807.

Pinto, J., & Shiffrar, M. (1999). Subconfigurations of the human form in the perception of biological motion displays. *Acta Psychologica*, *102*(2), 293–318.

Pollick, F. E., Lestou, V., Ryu, J., & Cho, S.-B. (2002). Estimating the efficiency of recognizing gender and affect from biological motion. *Vision Research*, *42*(20), 2345–2355.

Redcay, E., & Carlson, T. A. (2015). Rapid neural discrimination of communicative gestures. *Social Cognitive and Affective Neuroscience*, *10*(4), 545–551.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., & van Gerven, M. A. J. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, *180*, 253–266.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *ArXiv Preprint ArXiv:1406.2199*.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *ArXiv Preprint ArXiv:1212.0402*.

Theusner, S., de Lussanet, M. H. E., & Lappe, M. (2011). Adaptation to biological motion leads to a motion and a form aftereffect. *Attention, Perception, & Psychophysics*, *73*(6), 1843–1855.

Thurman, S. M., & Lu, H. (2014). Bayesian integration of position and orientation cues in perception of biological and non-biological forms. *Frontiers in Human Neuroscience*, *8*, 91.

Thurman, S. M., & Lu, H. (2016). Revisiting the importance of common body motion in human action perception. *Attention, Perception, & Psychophysics*, *78*(1), 30–36.

Thurman, S. M., van Boxtel, J. J. A., Monti, M. M., Chiang, J. N., & Lu, H. (2016). Neural adaptation in pSTS correlates with perceptual aftereffects to biological motion and with autistic traits. *NeuroImage*, *136*, 149–161.

Vaina, L. M., Solomon, J., Chowdhury, S., Sinha, P., & Belliveau, J. W. (2001). Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences*, *98*(20), 11656–11661.

van Boxtel, J. J. A., & Lu, H. (2013). A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. *Journal of Vision*, *13*(12), 7–7.

van Boxtel, J. J. A., & Lu, H. (2015). Joints and their relations as critical features in action discrimination: Evidence from a classification image method. *Journal of Vision*, *15*(1), 20–20.

Watson, J. D. G., Myers, R., Frackowiak, R. S. J., Hajnal, J. V., Woods, R. P., Mazziotta, J. C., Shipp, S., & Zeki, S. (1993). Area V5 of the Human Brain: Evidence from a Combined Study Using Positron Emission Tomography and Magnetic Resonance Imaging. *Cerebral Cortex*, *3*(2), 79–94.

Yamins, D., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). *Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream*.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.