Hierarchy in Biological and Computer Systems: A Review

Sawan Patel

Abstract

The presence of hierarchical organization in primate vision and motor function is a long-standing concept in neuroscience and has faced its fair share of skepticism. Though a theory describing a hierarchical organization for function has yet to be agreed upon, a hierarchy is anatomically agreed upon. Contention largely concerns the specific purpose of feedback signaling from higher cortical structures (IT, PFC, etc.) to lower cortical structures. Work has been done to understand the effects of feedback in a variety of stimulus-response experiments, but the precise nature of stimulus representations across the visual hierarchy is also debated when factoring the effect of context. The anatomical organization of the visual and motor processing pathways have been adapted by models in computer vision and robotics, respectively. In particular, this development has revolutionized object recognition tasks in computer vision. Functionally, however, these organizations in alternate disciplines also digress from the most recent discourse in neuroscience.

1. Introduction

The idea of a hierarchical organization of the visual processing pathway first emerged following Hubel and Wiesel's studies of the visual cortex in cats. Generally, they outlined the notion that receptive field properties vary in a patterned way as one goes up the visual cortex's laminar organization. Here, the receptive fields of neurons at one level of the envisioned hierarchy are constructed by combining the inputs from neurons at the immediately preceding level. After subsequent stages, small receptive fields tuned to simple stimuli are combined to form larger receptive fields tuned to more complex stimuli.

In the passing decades, more evidence has accumulated suggesting that a strict view of a visual hierarchy, as defined by Hubel and Wiesel, is exaggerated. However, the notion of a loose 'hierarchy' is still considered accurate. For instance, [36] describes how visual areas (such as V1) have several feed-forward projections to other areas (V2, V3, MT, etc.). Additionally, the processing pathways do adhere to several organizational principles that provide a framework for constructing an ordered scheme of organization. For instance, connections between cortical areas are organized in a reciprocal fashion (e.g. $(A \to B)$ and $(B \to A)$ for two distinct cortical areas A, B). Additionally, within a reciprocal pair of connections, there are characteristic differences in the laminar distributions of axonal terminations and cells of origin. In other words, projections originating from superficial layers and terminate in the granular layer (IV) are *forward* projections and, consequently, projections that originate in either superficial or deep layers that terminate outside of IV (most commonly in V1 or V2) are *backward* projections.

The notion of hierarchy is also evidenced to exist in a functional sense as well, where successively higher levels of hierarchy are associated with more advanced levels of visual analysis. It is also the generally the case that different visual areas at the same hierarchical level are involved in qualitatively different kinds of processing.

This idea of a hierarchical organization spread to computer vision through the 1980's and 1990's with the development of the neocognitron, the basis for convolutional neural networks, and the field has since continued developing its own models to optimize performance. These optimizations have, in some cases, deviated from findings in neuroscience, but several existing models continue to adapt components of the primate visual processing pathway as new information is uncovered. Hierarchical organization also has a history in robotics, particularly related to the organization of the primate motor system pairing of the central nervous system (brain, spinal cord) and the peripheries. In this review, I summarize recent empirical findings related to the anatomical and functional structure of primate vision in addition to patterns connecting hierarchically-organized models in computer vision and robotics. Future work relating the three fields by a more unified definition of hierarchy could not only guide further research in primate vision, but also develop more holistic and robust computer vision and robotics models.

2. Biological Mechanisms: Anatomy, Empirical Studies and Models

Here, I will survey across anatomical, electrophysiological and theoretical studies of the visual processing pathway in primate vision. These works will encompass current literature regarding the function of feedback signals and the dynamic representations of objects in the visual processing pathway. Across both sections, I aim to summarize the current answer to the tradeoff between translation invariance and progressive increase in complexity of features in the cortical hierarchy.

2.1. Feedback Signals

It is possible to establish a topological, hierarchical ordering of cortical areas purely based on feed-forward and feedback connectivity. [17], through quantitative analysis of the documented connectivity, produced an indeterminate solution to an exact hierarchical ordering of the primate vision network with 150,000 equally possible solutions given the known anatomical constraints. This was due to the lack of a measure for the hierarchical distance between a pair of cortical areas. Interestingly, quantifying the percentage of supragranular labeled neurons (SLN), which is a quantitative measure of the laminar distribution of parent neurons of cortical projections, allows for an deterministic solution to the connectivity problem irrespective of where a retrograde tracer is injected in the visual processing pathway [24]. SLN describes the percentage of parent neurons in a labeled area that are from the supragranular layer, following from the idea that feed-forward axonal projections originate in superficial layers of the cortex and terminate at the granular layer. Another study in mice similarly examined anterogradely labeled interareal projection patterns to match the visual areas to a consistent sequence of five overlapping hierarchical levels [10]. This was done by comparison of a novel metric, the optical density ratio (ODR), which compares the optical density of labeled axons in layers 2-4 to those in layers 1 added to layers 2-4, for each connection. This metric is sensible as feed-back connection terminations are much more frequent in layer 1 than in layers 2-4, whereas terminations in layers 2-4 are very frequent in ascending projections, making it a more accurate metric in comparison to SLN. From a metric perspective, describing the primate visual processing pathway, with feedback connections, as a hierarchy is more apt than not.

2.1.1 Counterstream Theory

Though it is commonly thought that feed-forward signaling generates receptive field properties and that feedback streams have a modulatory role, this notion conflicts with findings that characteristic physiological activity in higher areas being found in early visual cortices. Therefore, there is no operationally simple conception of ’higher’ and ’lower’ visual areas. It is more likely that the activation of feed-forward pathways instead give rise to rapid automatic characterization with little perceptual detail. Empty percepts are later ’supplied’ by the engagement of feedback pathways. Still, the feed-forward pathway is topologically organized in contrast to the more diffuse presence of feedback connections, which also are more numerous and evident across levels of the hierarchy.

The structural asymmetries between feed-forward and feed-back streams have led to the notion of a generative model, such that the prediction errors ascending the hierarchy and predictions descending the hierarchy reiteratively *interact*. This idea was originally conceived through the foundational interareal **counterstream** theory, where these two streams are segregated and converge at an area to interact with the local processing in the cortex [35]. This theory makes several predictions, but conflicts with studies showing that feed-back and feed-forward streams both involve varying proportions of cells in both supragranular and infragranular layers (i.e. the two pathways overlap) [4] and that pyramidal neurons projecting to lower cortical areas should not possess axon collaterals projecting to higher cortical areas (i.e. bi-directional connectivity) [35].

Novel studies follow the original counterstream theory with a proposal that, in a bayesian sense, the brain utilizes a generative model of the environment that explains ambiguous sensory information and simultaneously predicts future events. A trivial example is that an object’s distance to an observer is lost when light is projected onto the retina, so an inference has to be made given the two-dimensional disparity of object elements in both retinas to make a best estimate of such information. More specifically, feed-back signaling computes expectations about the incoming sensory stimuli to progressively lower levels of the cortex while feed-forward signaling computes the prediction error, which is propagated through ascending levels of the hierarchy. Notably, evidence supporting the notion that feed-forward signaling computes a prediction is plentiful, namely that when presented a predictable stimuli, the feed-forward pathway does not generate spikes, perhaps to save on energy costs [1].

It has also been argued that this model implies the cognitive penetration of vision, or that the generative models underlying prediction can affect visual perception [27]. These predictions are derived from cognitive, affective and contextual associations that provide important information which fills in the gaps left by raw sensory information. The balance between these bottom-up and top-down signals ought to be balanced in some way, which has been modeled previously using an adjustable learning rate parameter that determines the degree to which the prior expectations of a perception affect the current perception [18]. Evidence of such a tuning in primate vision has yet to be found.

2.1.2 Receptive field malleability

Another re-emerging foundational idea is that receptive field properties are subject to top-down influences, the nature of information conveyed by reentrant pathways, and how the information carried by neurons depends on behavioral context. Notably, over longer time periods, receptive fields change to accommodate alterations in visual experience, indicating that receptive fields have contextual influences and are more dynamic than previously thought. This idea draws back to an older work [6] showing that neurons in visual cortical areas can show selectivity for complex stimulus configurations. Here, a simple stimulus *outside* of the minimum response field for a parafoveal V1 cell can have a great facilitatory or inhibitory affect a neuron's response when presented jointly with a stimulus in the center of that cell's receptive field. It is clear that neuron responses are as dependent on characteristics of global contours and surfaces as they are to attributes of local features within their minimum response fields. Stimulus selectivity here can be determined through measuring tuning curves or mutual information (predicting stimulus identity (in bits) given a neuron's response).

Other examples of contextual influences on neural selectivity are evidenced through attention [11]. In one study, MT neurons, which characteristically respond to moving stimuli, were associatively trained to respond well to stationary stimuli given attentional cues. This suggests that activity is not just reflective of an external stimulus but also of cognitive state and stimulus associations. Additionally, frontal eye fields retain 'memory responses' in the absence of a visual stimulus while simultaneously representing the locations of intended saccades.

2.1.3 Grouping as an effect of Feedback

Studies show an apparent lack of computation done in a 'forward-pass' through the ventral visual pathway, such as how calculations done on neural conduction velocity demonstrate that there are only one or two spikes per cortical area before a decision is made when performing a classification task. This argues strongly against feed-back processing [34]. However, experiments including a vernier stimulus demonstrate that figure-level characteristics can affect basic feature processing, implicating that higher level representations have an effect on feature-level representations. They describe the importance of grouping at the level of our visual field, such that without it, human object recognition cannot be understood. Grouping, in the Gestalt context, pertains to the relating of particular elements to an object. In this case, grouping for complex stimuli highlights a fault in the purely feed-forward model. A similar study showed that grouping is time-sensitive, resulting in improved performance when a vernier-like stimulus surrounded by objects is presented for a longer amount of time [23]. Recent electrophysiological evidence indicates that the additional amount of time could allow for feed-back connections from the LOC to earlier cortical areas, ironing out perceptual grouping and facilitating improved classification.

These examples, and the aforementioned sections, describe how visual hierarchy is not simultaneously feed-forward and hierarchical in a strict sense. The ordering of areas from an organizational perspective is more anatomically correct than not, but functional differences when presenting a diverse collection of stimuli, from simple (e.g. parallel lines) to complex (faces, complex vernier stimuli, etc.), are quite apparent.

2.2. Object representations in visual processing pathway

A natural question to consider is how exactly are objects represented throughout the visual processing stream, as the evaluation of evolving representations across the cortical areas might yield evidence suggesting a hierarchical organization. However, the way objects themselves are represented in higher-order visual areas has been a point of controversy, despite the large number of experimental studies. Centrally, the issue to consider is to what degree are objects represented holistically in a manner that captures their global, configurational aspects and to what degree are they represented by their parts?

2.2.1 Representations in IT

This question has been studied with particular focus to area IT. One study recorded single neurons in the macaque IT supporting both representations and found that a large subset of neurons are sensitive to moderately complex features [33] while another subset is sensitive to more holistic representations, e.g. entire faces [12]. Interestingly, the results from studies similar to the above in anterior and lateral areas made it more possible to characterize the nature of object representations via common neuroimaging

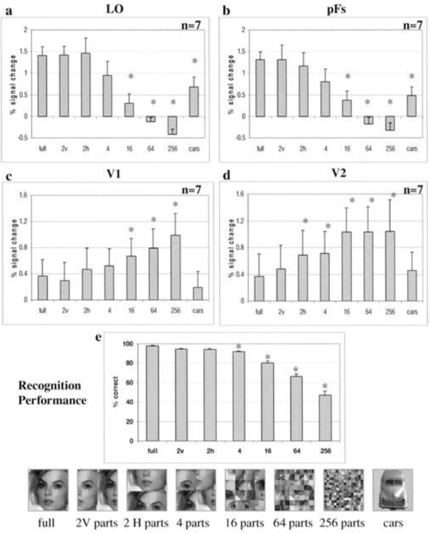


Figure 1. Activation across visual cortical areas for different stimulus configurations (full   entire image, 256   full image scrambled into 256 blocks, etc. [21]

methods, such as magentic resonance imaging (MRI). For example, a decade later, it was found that the majority of voxels in non-retinotopic object areas located in the lateral occipital lobe (LOC) remain active when images of a variety of objects are broken into scrambled blocks [16]. Such studies indicate feature-specific activation in earlier cortical areas prior to IT. A later study confirmed that this increasing sensitivity to feature complexity is largely consistent across the anterior-posterior axis, beginning in early retinotopic areas and proceeding into LOC [21]. With an MRI-based method, the investigators charted the time courses of activation obtained from regions showing the highest level of scrambling selectivity with early retinotopic areas to outline the drastically changing functional profile of the ventral visual stream in terms of sensitivity to image scrambling. Notably, this trend also held for non-facial stimuli, such as cars. However, the overlap between regions showing selectivity for face scrambling and car scrambling was not precise, which does not rule out the possibility of *multiple* hierarchical streams. The consistency of areas like LOC in responding to different objects sharing similar low-to-intermediate level features supports the existence of *a* hierarchical scheme, nevertheless.

Time-evolution of responses to facial features In [31], it was shown that the integration of facial information tends to proceed from the eyes and moves down the face through a time-series analysis of the N170 ERP signal. Notably, the integration of facial information stops once a marker indicating a face stimulus's emotional state was indicated (e.g. eyes in a 'fear' classified facial stimulus or the mouth in a 'happy' classified facial stimulus). The latency of each ERP particularly depended on the vertical distance of expression-specific diagnostic information from the two eyes, such that the 'eyes' in fear would lead to an early ERP. This process is both automatic, progressing down the y-axis of the face plane and goal-directed, as the integration stops once the diagnostic features have been integrated. A follow-up study performed a similar experiment where, on each trial, a subset of *action units* were selected and combined to produce an arbitrary 3D facial animation illustrated through four successive snapshots [19]. A subject was tasked with categorizing the random facial animation according to 6 possible emotions if the evolution of the facial stimulus correlated with their own subjective perceptual expectations of one of the particular emotions. They showed that the action units that were systematically expected (through analysis) early in the signaling dynamics comprise 'biologically-adaptive' action units, or those which would prompt the execution of an action that would be evolutionarily favorable (e.g. witnessing a 'jaw drop' prompted a perception of fear/confusion). The units that were expected later on in the dynamics comprise their diagnostic for categorizing the six classical emotions [13]. This reflects an evolving hierarchy of information over time, where the initial expressions elicit systematic confusions before later supporting accurate categorizations of the six categories of emotion. More generally, the prediction of the timing for signal transmission could present an adaptive advantage for autonomic responses (e.g. 'fight or flight'). In the wild, processing certain facial signals that yield some form of critical information earlier on could result in quicker decision-making in potentially life-threatening situations. For instance, these facial signals could have evolved as rapid behaviors to enhance sensory advantages (e.g. rapid muscle contractions protecting the eyes, or wrinkling the nose and mouth to reject noxious smells).

Context-dependence in IT As described in the vernier stimulus examples, context plays a critical role in visual object recognition. Notably, the regularity of our experiences of objects within environments and with each other can provide a rich environment that influences our recognition of objects and their categories. For instance, a face itself is embedded in a contextually rich environment, as it is accompanied by two ears and is attached via neck to a body. It has generally been found that neurons critical for face perception do not respond to these accompanying features. However, IT studies evaluate face-selectivity by presenting faces *in isolation* rather than as they typically appear. In studies where the activations of cortical regions has been examined when considering contextually-supported faces, it has been shown that face-selective IT neurons respond not only to images of faces, but to parts of an image where contextual clues indicate a face *ought* to be, even in its absence. [3] recorded simultaneously from middle lateral and posterior lateral 'face-selective' patches of neurons in rhesus monkeys. In one experiment, they presented complex scenes that lacked faces but contained cues indicating where a face ought to be, such as a body. The middle lateral face cells responded to the parts of the images where the face *would have* been located, even when the faces were occluded and not just absent. However, the latency of response was slower than responses for intact faces by 30ms. Another experiment entailed presenting images where a non-face object was present in an image twice, once above a body and once not. Responses to the non-face object positioned above a body were larger than responses to the same object when not positioned above a body, both at the population level and in individual channels. They also found that responses to faces above bodies were *indistinguishable* from responses to disembodied faces, indicating that the presence of bodies within the receptive field of face-selective neurons only provides a supporting role. The above illustrates how IT neurons do not code objects and complex shapes in isolation. Rather, evidence suggests that these neurons are sensitive to statistical regularities of cumulative experience. The facilitation of face-selective neurons when a clear face is not even present reflects a lifetime of experience where bodies are usually accompanied by faces. The retinotopic spatial regularity between bodies and faces might explain why face and body 'patches' often emerge in adjacent parts of the visual cortex. Interestingly, the group reported that the posterior lateral responses always preceded the middle lateral responses, in both the clear face and occluded face conditions. Given the anatomical organization of the two patches, it is unlikely that the described 'body-facilitation' is an effect of feedback signaling.

2.2.2 Abstract representations of faces

Further concerning area IT, what about abstract facial stimuli? That is, to what degree does the highest level of the ventral visual pathway distinguish between face-like stimuli and actual faces, if at all? Face pareidolia, the attribution of real face traits to non-face objects due to illusory perceptions, can erroneously activate a connection between visual input areas and internal representations of faces. Naturally, the fusiform-face area (FFA) plays a crucial role in the perception of both actual faces and illusory face perceptions. In an fMRI study where several volunteers were presented grayscale instances of real faces, face pareidolia and scrambled images, both real face and face-pereidolia specific activation was found across FFA, prefrontal cortex (PFC) and V1/V2 [2]. The asynchronous conjunction of PFC and early occipitotemporal activity suggests that there is some coordination of bottom-up and top-down processing in either case. It has also been suggested that the bilateral activation of FFA could result from the *expectation* of seeing faces rather than from the actual face pareidolia perception [14]. The procedure did not allow for an accurate determination of the temporal ordering of activations across different areas.

Pareidolia provides an interesting set of stimuli to test the counterstream theory, as the bilateral activation of the FFA could reflect a forward-pass 'correction' of the expectation that a face stimulus is going to be presented. This feed-back signaling, perhaps originating in PFC and descending to the FFA, is potentially overwritten by the feed-forward confirmation through raw sensory information that the stimulus is not in fact a face. An analysis of the temporal dynamics through a more powerful fMRI method could clarify the precise ordering of activation across the upper-cortical regions in the visual processing pathway.

3. Models in Computer Vision and Robotics

Older models in computer vision for object recognition drew from the ideas of David Marr, that what we see is a fully elaborated diagram of a visual scene by a transformation from two-dimensional input data projected onto the retina into a three-dimensional spatio-temporal model. These ideas connected directly to Hubel and Wiesel's strict, feed-forward hierarchical processing proposal. However, recent models have begun to relax these assumptions, in light of the prevalence of back-propagations, the importance of attention in guiding visuo-motor collaboration and integrating motion information to simplify computation in recognition tasks.

3.1. Overview of Models

Optical Flow In [22], the authors demonstrate a novel technique (appearance-motion decomposition, AMD) for zero-shot segmentation on novel images by *segment flow* constructed from a a motion network output for two consecutive frames and a segmentation network that produces a set of relevant object masks. These segment flows are warped into a predicted image of the second of the consecutive images, and the model is trained to minimize the error of the prediction and the actual image. Though this method deviates from strategies thought to be used by humans for segmentation (i.e. processing of features with progressively increasing complexity), it could be extended to implement an attentional mechanism with a focus on dynamically changing areas between successive frames when presented with a collection of frames in a video instead of just two. A similar work has since combined the strengths of motion-based and appearance-based segmentation by supervising an image segmentation network with the task of predicting regions likely to harbor simple motion patterns [9]. The use of motion data for supervision allows for segmentation in videos and in still images without the need for manual annotation.

Attention-Dependent Models [8] leverages a novel attention mechanism for fixation, a retina-inspired approach to pixel sampling, and a network architecture inspired by the ventral visual stream to produce a model that is resistant to commonly-found adversarial images/classes in typical classification approaches. This approach not only generalizes performance to possible adversarial examples with adaptive eye movement, but presents a method for augmenting a training set via retinal sampling, which naturally results in a selection of distorted views from a particular input image. This method could be integrated into any existing model architecture with little difficulty following a modification of the described ventral pathway. The exact architecture described in this paper resulted in sub-par performance compared to similar architecture using a different pixel sampling method, which they attribute to the differences between images contained in object recognition datasets (e.g. ImageNet) and naturalistic images. For example, objects of interest in naturalistic images are often a fragment of the entire visual field, whereas the target object in the majority of object recognition datasets takes up the majority of the image.

Recurrency The presence of recurrency in human object recognition is well-documented, with recent work investigating the temporal dynamics of signaling during modified object recognition tasks [26, 30, 37]. Appropriately, papers in computer vision have implemented recurrency across several architectures. [32] describes a series of network architectures that incorporate any combination of feed-forward, lateral and recurrent connections which are tested against modifications of the MNIST dataset, one of which tests model resistance to a form of occlusion. The most complex model described features, at any one layer, lateral connections between all units in one layer, recurrent connections to the previous layer and feed-forward connections to the subsequent layer, all of which are implemented via deconvolution/convolution. This particular architecture demonstrated the best performance across all developed architectures, including performance on the debris occlusion data. If the desire is to explicitly replicate biological signaling and performance, [20] describes the deployment of CORnet-S, a shallow four-layer network that held the largest Brain-Score (a benchmark composed of neural recordings and behavioral measurements) of all submitted base models. Each layer in this network is anatomically mapped to a critical region in primate ventral visual pathway (V1, V2, V4, IT), and it terminates with a linear decoder that maps the network output to the output behavioral choices. Each layer is mapped to the subsequent layer via convolution, with kernels of differing parameters. Recurrence is implemented by passing the output of one particular layer *back* to itself several times. The Brain-Score metric specifically measures how well models can predict (a) mean neural response for each neural recording site to each tested naturalistic image in non-human primate V4 and IT, (b) mean-pooled human choices when reporting a target object in each naturalistic image, and (c) when object category is resolved in non-human primate IT. Performance was compared across a variety of architectures (e.g. AlexNet, ResNet, DenseNet, MobileNet, Inception, etc.), and CORnet-S achieved the highest performance. It should be understood that the direct comparison of network responses to mean neural responses for each input may not be a relevant factor in the pure task of object recognition, however. Since, several architectures (many of which are built off of ResNet) have surpassed CORnet-S.

Summary The use of motion information for supervision aligns with the counterstream theory's proposed dichotomy of the visual processing pathway, as the brain areas harboring neurons sensitive to optic flow are higher cortical regions (e.g. V2, V3A, V4, MT, etc.) that are outputs for feed-forward signaling. Each of these regions shows a degree of functional sensitivity by motion type (e.g. local vs global, self vs object motion, radial vs rotational vs translational flow, etc.). Additionally, the simulation of attention mechanisms by sampling fixation points with the highest saliency to direct focus onto particular patches of an input image directly mirrors the saccadic mechanisms controlled by the superior colliculus and frontal eye fields. This is also paired with graded retinal sampling to avoid the stereotypically flat sensory input characteristic of CNNs. Another welcome progression is the implementation of recurrency in typical recognition models of various designs. In many cases, the degree of recurrency pales in comparison to that found in primate visual anatomy but certainly demonstrate more than adequate performance. Surveying across all models shown on Brain-Score demonstrates the importance of recurrency, though these models do not strictly boast the same performance gains as models designed without precisely mirroring neural recordings in mind.

3.2. Pooling mechanisms

The precise mechanism for pooling in the primate visual recognition pathway is unclear. However, advances in pooling mechanisms in computer vision provide several benefits over the simple max-pooling or mean-pooling mechanisms as information loss is apparent in either case. [15] proposes multi-scale order-less pooling (MOP), which extracts local patches at a single scale and then pools them over regions of increasing scale. The output following the extraction of patches of each feature map results in one map of the entire image. [5] demonstrates a genetic pooling algorithm. Here, a population of attentional weights are generated randomly between the interval [0, 1] in the first generation. The model is trained for each set of attention weights in the population and error is calculated through the corresponding loss functions. Then through generations, these attention weights are optimized to achieve minimum loss.

3.3. Representation learning

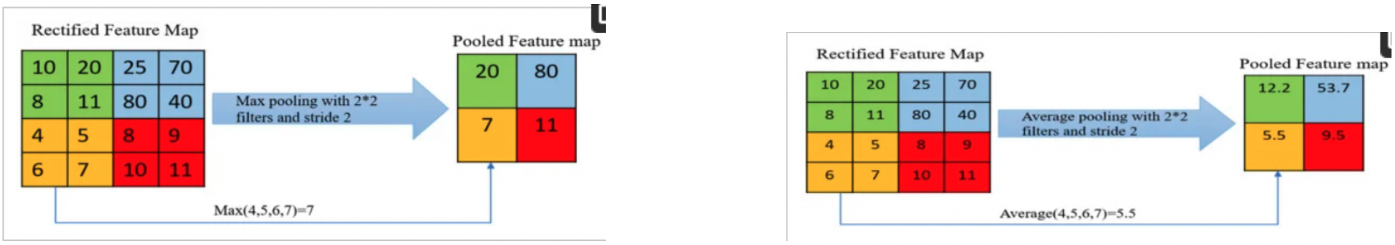
Another application of visual hierarchy can be found in representation learning. A good representation is typically characterized as capturing multiple configurations from the input and can also organize the explanatory factors of the input into a hierarchy (with more abstract concepts at the top). Many deep neural networks fail at learning reliable representations as they are heavily dependent on the training objective (e.g. may focus on shapes and parts of objects in an object recognition task but not rotation). The discriminative features learned from solving high-level image classification tasks might not be appropriate for mid and low-level tasks, which reduces their transferability. Generative adversarial networks (GANs) have been found to encode rich hierarchical semantics in layer-wise representations, but are limited in that they are designed for image generation and not inference (e.g. taking an image and extracting its features). In [38], the authors show that a pre-trained GAN can be considered as a learned loss function, which is combined with a novel hierarchical encoder whose outputs align with the layer-wise representations of the generator. The generator therefore takes the feature hierarchy produced by the novel encoder as per-layer inputs and supervises the encoder by reconstructing the original input image into features called *generative hierarchical features*. This methodology can be applied to a wide range of discrimination tasks, achieving near-top performance on digit recognition tasks, face verification and on ImageNet. They find that features at lower levels are more suitable for lower level tasks (e.g. luminance regression) and those at higher levels are better suited for higher-level tasks (e.g. pose estimation).

3.4. Motor hierarchy

Hierarchical processing is also evident in robotics. The task is compared to reaching in a biological organism, whereby the nervous system integrates different sensory modalities and coordinate multiple degrees of freedom in the human arm to achieve the feat. However, many challenges are present: the noise and transport delays in neural signals, fatigable muscles and unpredictable environmental disturbances. Despite this, the task can still be accomplished. Notably, there is a hierarchical organization of neural structures underlying *movement*, where each layer performs a specific function that increases in abstraction. Additionally, there is evidence that there are independent 'lower levels' of this organization (e.g. spinal cord) capable of relatively complex motor behaviors independent of the higher levels, such as cats with transected spinal cords capable of learning to walk on a treadmill. This proposed hierarchical paradigm is therefore accompanied with many questions, such as each level's function, their limits and if the existence of a global hierarchy is even accurate.

The idea of hierarchical architectures in robotics has existed since the proposal of subsumption architecture [7]. Here, a collection of layers that each specify a behavior pattern for a robot each consisting of a network of 'message passing augmented finite state machines.' Each augmented FSM has a set of registers and a set of synchronized timers connected to a typical FSM that can control a combinatorial network fed by the registers. The arrival of a message to the network can trigger a change in the interior FSM. This machine can be further augmented by adding new machines that can provide input to each other via registers capable of inhibition or excitation. Robots incorporating this type of model are capable of walking, load balancing, and several other behaviors.

Recently, hierarchical architectures currently allow for the modularization and simplification of individual controllers and training procedures. Each subsystem deals with a fraction of the incoming sensory information and can be trained separately with its own distinct cost functions and performance requirements. These sorts of models are far more flexible than 'flat' non-hierarchical controllers that simultaneously process all sensory information and directly calculate behavioral output. In [25], the authors explore a series of distinct, hierarchically organized models of arm control built of multiple cascading layers of simple proportional and proportional-derivative feedback loops with low-pass filtering [28, 29]. This work adapts such proposed hierarchical control architecture to a 4 DOF robot arm to explore its theoretical capabilities in deployment. They demonstrate several fundamental invariant properties found in human hand trajectories, such as isochrony, bell-



(a) Information loss from a max-pooling motif when inputs in the sampled space are roughly equivalent to maximum value.

(b) Information loss from a mean-pooling motif when inputs in the sampled space share a high degree of variance.

shaped velocity profiles, and the speed-curvature power law, which are also found in the robot arm trajectories without planning or optimization. This architecture is also able to spontaneously adapt in behavior whenever its wrist joint is blocked, if the visual field is rotated with respect to the arm, and when the robot hand is extended by a tool.

4. Conclusion

Though all of the aforementioned models, save for [8], deviate greatly from the most well-respected models of primate visual hierarchy, they each incorporate unique elements of counterstream theory and modularity. Primate vision is still not fully understood, most notably the precise mechanism underlying the tuning of feedback influences on visual perception. However, the general notion of a hierarchical organization across vision and motor systems produces adaptive gains across several tasks and provides the flexibility to design independent modules for processing across a variety of different tasks, many of which are reused and many of which are not required depending on the task. A promising area of future research in neurophysiology that would undoubtedly benefit computer vision is the precise pooling mechanisms utilized from one 'level' of a the hierarchy to the next for the ascending pathway. Though spatial resolution across several areas is a challenge for current methods that allow for the temporal resolution required to measure representations at the level of spikes, further work in this area could both confirm the near-serial processing in the prediction-error computation mechanism and provide models in computer vision a novel pooling mechanism that has been tested by evolution.

References

[1] Laurence Aitchison and Máté Lengyel. With or without you: predictive coding and bayesian inference in the brain. *Current opinion in neurobiology*, 46:219–227, 2017. 2

[2] Gulsum Akdeniz, Sila Toker, and Ibrahim Atli. Neural mechanisms underlying visual pareidolia processing: an fmri study. *Pakistan Journal of Medical Sciences*, 34(6):1560, 2018. 5

[3] Michael J Arcaro, Carlos Ponce, and Margaret Livingstone. The neurons that mistook a hat for a face. *Elife*, 9:e53798, 2020. 5

[4] Pascal Barone, Alexandre Batardiere, Kenneth Knoblauch, and Henry Kennedy. Laminar distribution of neurons in extrastriate areas projecting to visual areas v1 and v4 correlates with the hierarchical rank and indicates the operation of a distance rule. *Journal of Neuroscience*, 20(9):3263–3281, 2000. 2

[5] Kamanasish Bhattacharjee, Millie Pant, Yu-Dong Zhang, and Suresh Chandra Satapathy. Multiple instance learning with genetic pooling for medical data analysis. *Pattern Recognition Letters*, 133:247–255, 2020. 7

[6] PO Bishop. Neural mechanisms for binocular depth discrimination. In *Sensory Functions*, pages 441–449. Elsevier, 1981. 3

[7] Rodney A Brooks. A robot that walks; emergent behaviors from a carefully evolved network. *Neural computation*, 1(2):253–262, 1989. 7

[8] Minkyu Choi, Yizhen Zhang, Kuan Han, Xiaokai Wang, and Zhongming Liu. Human eyes inspired recurrent neural networks are more robust against adversarial noises. *arXiv preprint arXiv:2206.07282*, 2022. 6, 8

[9] Subhabrata Choudhury, Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Guess what moves: Unsupervised video and image segmentation by anticipating motion, 2022. 6

[10] D'Souza Rinaldo D, Wang Quanxin, Ji Weiqing, Meier Andrew M, Kennedy Henry, Knoblauch Kenneth, and Burkhalter Andreas. Hierarchical and nonhierarchical features of the mouse visual cortical network. *Nature communications*, 13(1):503, 2022. 2

[11] Gilbert Charles D and Li Wu. Top-down influences on visual processing. *Nature Reviews Neuroscience*, 14(5):350–363, 2013. 3

[12] Robert Desimone. Face-selective cells in the temporal cortex of monkeys. *Journal of cognitive neuroscience*, 3(1):1–8, 1991. 3

[13] AK Anderson DH Lee, JM Susskind. Social transmission of the sensory benefits of eye widening in fear expressions. *Psychological Science*, 24:957–965, 2013. 4

[14] Michael Esterman, Benjamin J Tamber-Rosenau, Yu-Chin Chiu, and Steven Yantis. Avoiding non-independence in fmri data analysis: leave one subject out. *Neuroimage*, 50(2):572–576, 2010. 5

[15] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. *CoRR*, abs/1403.1840, 2014. 7

[16] Kalanit Grill-Spector, Tammar Kushnir, Talma Hendler, Shimon Edelman, Yacov Itzchak, and Rafael Malach. A sequence of object-processing stages revealed by fmri in the human occipital lobe. *Human brain mapping*, 6(4):316–328, 1998. 4

[17] Claus-C Hilgetag, Mark A O'Neill, and Malcolm P Young. Indeterminate organization of the visual system. *Science*, 271(5250):776–777, 1996. 2

[18] J Hohwy. Priors in perception: Top-down modulation. *Bayesian perceptual learning*, 2017. 3

[19] Rachael E Jack, Oliver GB Garrod, and Philippe G Schyns. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192, 2014. 4

[20] Jonas Kubilius, Martin Schrimpf, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo. Brain-like object recognition with high-performing shallow recurrent anns. *CoRR*, abs/1909.06161, 2019. 6

[21] Yulia Lerner, Talma Hendler, Dafna Ben-Bashat, Michal Harel, and Rafael Malach. A hierarchical axis of object processing stages in the human visual cortex. *Cerebral cortex*, 11(4):287–297, 2001. 4

[22] Runtao Liu, Zhirong Wu, Stella Yu, and Stephen Lin. The emergence of objectness: Learning zero-shot segmentation from videos. *Advances in Neural Information Processing Systems*, 34:13137–13152, 2021. 5

[23] Mauro Manassi, Bilge Sayim, and Michael H Herzog. When crowding of crowding leads to uncrowding. *Journal of vision*, 13(13):10–10, 2013. 3

[24] Nikola T Markov, Julien Vezoli, Pascal Chameau, Arnaud Falchier, René Quilodran, Cyril Huissoud, Camille Lamy, Pierre Misery, Pascale Giroud, Shimon Ullman, et al. Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology*, 522(1):225–259, 2014. 2

[25] Adam Matić, Pavle Valerjev, and Alex Gomez-Marin. Hierarchical control of visually-guided movements in a 3d-printed robot arm. *Frontiers in Neurorobotics*, page 149, 2021. 7

[26] Yalda Mohsenzadeh, Sheng Qin, Radoslaw M Cichy, and Dimitrios Pantazis. Ultra-rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway. *eLife*, 7:e36329, jun 2018. 6

[27] Claire O'Callaghan, Kestutis Kveraga, James M Shine, Reginald B Adams Jr, and Moshe Bar. Predictions penetrate perception: Converging insights from brain, behaviour and disorder. *Consciousness and cognition*, 47:63–74, 2017. 2

[28] William T Powers. A model of kinesthetically and visually controlled arm movement. *International journal of human-computer studies*, 50(6):463–479, 1999. 7

[29] William T Powers. Living control systems iii: The fact of control. 2008. 7

[30] Karim Rajaei, Yalda Mohsenzadeh, Reza Ebrahimpour, and Seyed-Mahdi Khaligh-Razavi. Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS computational biology*, 15(5):e1007001, 2019. 6

[31] Philippe G Schyns, Lucy S Petro, and Marie L Smith. Dynamics of visual information integration in the brain for categorizing facial expressions. *Current biology*, 17(18):1580–1585, 2007. 4

[32] Courtney J Spoerer, Patrick McClure, and Nikolaus Kriegeskorte. Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in psychology*, 8:1551, 2017. 6

[33] Keiji Tanaka, Hide-aki Saito, Yoshiro Fukada, and Madoka Moriya. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of neurophysiology*, 66(1):170–189, 1991. 3

[34] Simon Thorpe, Arnaud Delorme, and Rufin Van Rullen. Spike-based strategies for rapid processing. *Neural networks*, 14(6-7):715–725, 2001. 3

[35] Shimon Ullman. Sequence seeking and counter streams: a computational model for bidirectional information flow in the visual cortex. *Cerebral cortex*, 5(1):1–11, 1995. 2

[36] David C Van Essen and John HR Maunsell. Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, 6:370–375, 1983. 1

[37] Dean Wyatte, David J Jilk, and Randall C O'Reilly. Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in psychology*, 5:674, 2014. 6

[38] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4432–4442, 2021. 7