Université de Montréal

Évidences psychophysiques que l'information visuelle de bas niveau peut influencer les interférences en mémoire à long terme

Par
Jean-Maxime Larouche

Département de psychologie, Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de la maîtrise
en psychologie, option neuroscience cognitive

08/2020

Université de Montréal
Unité académique : département de psychologie, Faculté des arts et des sciences

Ce mémoire intitulé

Évidences psychophysiques que l'information visuelle de bas niveau peut influencer les interférences en mémoire à long terme

Présenté par

Jean-Maxime Larouche

A été évalué(e) par un jury composé des personnes suivantes

Greg West
Président-rapporteur

Frédéric Gosselin
Directeur de recherche

Pierre Bellec
Membre du jury

Karim Jerby
Membre du jury

Résumé

Ce mémoire démontre que les caractéristiques de niveau inférieur telles que les fréquences spatiales (SF) et les orientations (V1) sont encodées spécifiquement avec les caractéristiques visuelles de niveau supérieur auxquelles elles ont été associées lors de l'apprentissage. Deux groupes ont appris deux ensembles de visages - composés soit de la même combinaison soit de deux combinaisons différentes de SF et d'orientations entre les ensembles. Plus tard, dans une tâche à trois alternatives, les participants ont distingué les visages appris des deux ensembles et des nouveaux ensembles de visages non cibles (filtrés avec trois filtres complémentaires de bas niveau). Il existe des preuves extrêmement solides que la similarité de bas niveau entre les deux ensembles appris augmente les interférences de la mémoire de reconnaissance (BF01 : 4,1; $p<0.01$) et il existe également des preuves solides que la similarité de bas niveau observée avec les nouvelles fonctionnalités de haut niveau n'a pas d'impact sur les interférences (BF01 : Inf; $p=1$). Nos résultats expliquent les contradictions apparentes dans la littérature, démontrent une directionnalité évidente dans le modèle d'encodage de la mémoire humaine et aident à consolider la relation entre l'esprit humain et les réseaux de neurones profonds.

Mots-clés : Mémoire à long-terme, Interférences, Cortex visuel primaire, Psychophysique, Apprentissage perceptuel

Abstract

This study demonstrates that lower level features like spatial-frequencies and orientations (V1) are encoded specifically with the higher level visual features they were associated with during learning. Two groups learned two sets of faces − composed either of the same or of two different combinations of SF and orientations between the sets. Later, in a task with three alternatives, the participants distinguished the learned faces from the two sets and new non-target face sets (filtered with three low-level complementary filters). There is extremely strong evidence that the low-level similarity between the two learned sets increases recognition memory interference (BF10: 4,1355; $p<0.01$) and there is also strong evidence that low-level similarity seen with new high-level features is not impacting interferences (BF01: Inf; $p=1$). Our findings explain apparent contradictions in the litterature, demonstrate an evident directionality in the human memory encoding model and help consolidate the relation between the human mind and deep neural networks.

Keywords : Long-term memory, Interference, Early visual cortex, Top-down, Psychophysics, Perceptual learning

Table des matières

Liste des figures

Liste des sigles et abréviations

NTI : non-target interferences; interferences des images non-cibles

PI : proactive interferences; interferences proactives

RI : retroactive interferences; interferences rétroactives

RHT : reverse hierarchy theory; théorie de la hiérarchie inversée

SF : spatial frequency; fréquences spatiales

Remerciements

nani gigantum humeris insidentes

« Des nains sur les épaules de géants » nous dit Bernard de Chartres. Puis en 1675, Newton reprend la métaphore : « Si j'ai vue plus loin, c'est en montant sur les épaules de géants ». Murray Gell-Mann à répliqué à l'obtention de son prix Nobel en 1969 : « Si j'ai vu plus loin que d'autres, c'est que je suis entouré de nains ». Être plus haut, c'est voir plus loin. Ce géant ou celui qui est plus grand, c'est l'évolution de notre connaissance, grandissant de génération en génération. De ce fait, je me dois ici souligner l'influence importante de plusieurs penseurs, auteurs, artistes, scientifiques et philosophes qui m'ont directement et indirectement permis de réaliser ce mémoire.

Plus spécifiquement, je tiens à remercier mon directeur de recherche Frédéric Gosselin sans qui le projet n'aurait pas pu avoir lieu. Plus personnellement, je remercie ma famille, mes amis, mes collègues et surtout ma fiancée Clémentine Pagès. Autant pour ton soutien émotionnel, physique qu'intellectuel, merci d'être et d'exister. Tu es du sel pour ma terre, donnant du goût à ma vie.

Chapitre 1 – Introduction

« Il avait appris sans effort l'anglais, le français, le portugais, le latin. Je soupçonne cependant qu'il n'était pas très capable de penser. Penser c'est oublier des différences, c'est généraliser, abstraire. Dans le monde surchargé de Funes, il n'y avait que des détails, presque immédiats. »

- (Borges, 1957)

La particularité du personnage de Borges est qu'après son accident à cheval, Irénée Funes n'avait même plus besoin de penser pour se rappeler, sa mémoire était presque infinie. En fait, ses souvenirs étaient si distincts, qu'il ne pouvait plus comprendre comment le symbole générique du chien pouvait embrasser à la fois « le chien de trois heures quatorze (vu de profil) [et] le chien de trois heures un quart (vu de face) ». Pour nous, ayant une capacité mnémonique plus limitée que Funes, les catégorisations de nos perceptions − réduisant et classifiant l'information − constituent des indices fondamentaux nécessaires au rappel de nos souvenirs. En fait, suivant la combinaison d'un monde perceptif complexe et de notre capacité limitée, la pensée – celle qui généralise, catégorise et abstrait – joue un rôle important dans le contrôle des interférences entre nos différents souvenirs (Brady et al., 2011), nous protégeant du même fait de l'oubli (Crowder, 2014). En fait, les systèmes complexes ont fréquemment recours à une architecture hiérarchique, ayant des composantes qui sont indépendantes dans leur contenu spécifique. Que ce soit des hiérarchies physiques, biologiques, chimiques, sociales, etc. Sur le plan théorique, on peut s'attendre à ce que la complexité évolue en hiérarchie, car dans leur dynamique, les hiérarchies ont la propriété de simplifier l'information traitée par un système (Simon, 1991). La solution la plus simple étant souvent la mieux adaptée. C'est en ce sens que nos différents systèmes perceptuels ont évolué vers une architecture hiérarchique nous permettant aujourd'hui de simplifier et d'encoder l'information complexe de notre réalité.

Tentant de comprendre comment se structure l'encodage d'information en mémoire chez l'humain, les chercheurs en sciences cognitives ont découvert assez tôt que l'interférence entre deux souvenirs augmente en fonction de leur « similarité » de représentation, et ce, pour la sémantique (Osgood 1946, McGaugh 2000), pour les contextes d'encodage (Bilodeau et Schlosberg 1951), etc. Puis, en 1966, quelques études conduites par Baddeley ont démontré que les interférences en mémoire à long terme sont affectées par la similarité des représentations de haut niveau (sémantique des mots), mais − contrairement à la mémoire à court terme − ne sont pas affectées par la similarité des propriétés de bas niveau (acoustique des mots) (Baddeley, 1966a, 1966b). Plus tard, Konkle et al. (2010) ont démontré que la similarité de certaines catégories d'objets affecte davantage les interférences en mémoire à long terme. Ensuite, en mesurant la similarité de bas niveau séparément pour la couleur, la taille et l'apparence visuelle globale, ils ont également déduit que la similarité des représentations perceptuelles ne prédisait pas les interférences de mémoire à long terme. Ils proposèrent donc que les catégories sémantiques fournissent des « crochets conceptuels » qui permettent de récupérer les traces mnémoniques complètes, incluant non seulement les données sémantiques abstraites, mais également des informations perceptuelles sur les détails des objets (Konkle et al., 2010a, Konkle et al., 2010b). Par ailleurs, d'autres études sur les interférences en mémoire à long terme arrivent à des résultats similaires (p.ex. Ishai et Sagi 1995. 1997).

Cependant, toutes ces études utilisent des protocoles favorisant l'apprentissage associatif par le lobe temporal médian, en mémoire épisodique, parce qu'elles comportaient peu ou pas de répétition des stimuli présentés et ne récompensaient pas les réponses exactes des participants (Mahut et al., 1982; Wimmer et al., 2014). Or le cerveau humain possède d'autres systèmes

d'apprentissage en constante interaction, parfois en compétition (Packard et Knowlton 2002). Notamment, l'apprentissage perceptuel est un processus lent et dirigé statistiquement qui permet de faire face à la complexité du monde réel et des propriétés qui le composent (Karni et Sagi, 1993). En conséquence, plusieurs expositions aux stimuli sont nécessaires − en utilisant l'erreur de prédiction − afin d'encoder correctement les informations perceptuelles utiles pour une tâche demandée (Packard et Knowlton, 2002; Montague et al., 2004; Law et Gold, 2009, Wimmer et al., 2014). Concrètement, les régions cérébrales de bas niveau − permettant, par exemple, la perception visuelle (V1-V2) − démontrent une forme de plasticité neuronale résultant de retours informationnels dirigés soit par des représentations de plus haut niveau dans la hiérarchie du système visuel, soit par des récompenses provenant de régions externes acheminées entre autres par le striatum qui − contrairement au lobe temporal médian − reçoit et envoie des informations presque partout dans le cerveau, même à V1 (McGeorge et Faull, 1989; Petrov et al. 2005). Il devrait donc être possible de créer une association, médiée par le striatum, entre une réponse comportementale de haut niveau et des propriétés visuelles de bas niveau (e.g. fréquences et orientations spatiales dans V1) et ce, même si ces propriétés ne se distinguent pas consciemment dans les régions de niveau supérieur de la hiérarchie du traitement visuel (e.g. IT). Ainsi, le but premier de cette étude est de tester l'hypothèse selon laquelle les propriétés visuelles de bas niveau influencent les interférences en mémoire à long terme lorsque leur association avec une réponse comportementale de haut niveau a été récompensée à plusieurs reprises.

Le traitement sensoriel implique donc des connexions directes le long d'une hiérarchie d'aires corticales représentant des aspects progressivement plus complexes et abstraits de la scène visuelle. Superposées à ces voies sensorielles ascendantes, il existe aussi des connexions de

rétroaction qui transmettent des informations d'ordre supérieur aux représentations corticales antécédentes. En fait, les neurones corticaux sont soumis à des influences descendantes de l'attention, de tâche perceptuelle, des attentes, etc. (Gilbert et Li, 2013). Les propriétés fonctionnelles des neurones ne sont donc pas fixes; on pense maintenant qu'ils s'adaptent à aux influences de plus haut niveau (pour une revue, voir Gilbert et Li, 2013), modifiant leur fonction par rapport au contexte dans lequel ils se trouvent. Ces influences ont été observées à tous les stades de la hiérarchie visuelle, y compris dans V1 et même dans les corps genouillés latéraux du thalamus (Li et al., 2004; O'Connor et al., 2002), et ont un rôle important dans l'encodage et le rappel des informations apprises. Par exemple, dans le domaine de l'apprentissage perceptuel, Hochstein et Ahissar (2004) ont proposé la théorie de la hiérarchie inversée (RHT). Selon eux, l'apprentissage perceptuel s'effectue de manière descendante, en commençant à un niveau supérieur de traitement et en rétrogradant jusqu'au niveau d'entrée lorsqu'un meilleur rapport signale sur bruit est nécessaire pour effectuer une tâche. Toujours selon ces chercheurs, les neurones du cortex visuel primaire montrent une plus grande modulation aux changements de position des composantes pertinentes pour une tâche que des composantes non liées à la tâche (Ahissar and Hochtein, 2002, 2004). Aussi, ces propriétés neuronales associées à l'apprentissage perceptuel ne sont présentes que lorsque l'observateur accomplit la tâche pour laquelle il a été entraîné (Li et al., 2004; Li et al., 2008; McManus et al., 2011). Ces influences descendantes ont donc un rôle important sur l'apprentissage perceptuel, étant nécessaires à l'encodage de l'information ainsi qu'à son rappel. On peut ainsi s'attendre à ce que l'encodage des propriétés de bas niveau soit spécifique à la représentation de plus haut niveau à laquelle l'encodage de ces propriétés de bas niveau a été associé durant l'entraînement. Bref, le second but de cette étude est

de démontrer que pour rappeler une représentation spécifique de bas niveau, la bonne représentation de haut niveau − celle associée à la représentation de bas niveau durant l'entraînement − est nécessaire.

Dans un premier temps, cette recherche démontre que la similarité des propriétés visuelles de bas niveau (fréquences spatiales et orientations) influence les interférences en mémoire à long terme, lorsque ces propriétés sont utiles à la tâche demandée et sont apprises après plusieurs expositions aux stimuli. Dans un second temps, cette recherche corrobore l'hypothèse selon laquelle l'information de bas niveau (~V1) est seulement encodée de manière spécifique à sa représentation dans un niveau supérieur du traitement visuel (~IT) − celle ayant été associée à la représentation de bas niveau durant l'apprentissage à l'aide d'évidences psychophysiques. Nous conclurons que la mémoire à long terme structure l'encodage et le rappel des informations de manière descendante en priorisant des représentations plus générales et abstraites (de haut niveau) afin de récupérer des informations plus précises et spécifiques à des situations lorsque nécessaire (de bas niveau).

Contribution des auteurs à l'article

L'article du présent mémoire, intitulé « Psychophysical evidence that primary visual information can impact memory interference in long-term memory » a été rédigé par Jean-Maxime Larouche sous la supervision de Frédéric Gosselin. Ce manuscrit sera soumis pour publication dans une revue évaluée par les pairs. L'idée originale du projet (les quasi-métamères, les hypothèses et la procédure utilisée) vient de Jean-Maxime Larouche et de Frédéric Gosselin. Le code expérimental a été écrit principalement par Jean-Maxime Larouche. La mise en place de l'étude (ex. le recrutement/passation des participants) a été supervisée par Jean-Maxime

Larouche. Le code d'analyse, le traitement statistique des données, la simulation ainsi que les figures ont été créés par Jean-Maxime Larouche. La recension de la littérature a été effectuée par Jean-Maxime Larouche et l'écriture de l'article a été faite par Jean-Maxime Larouche et Frédéric Gosselin.

Chapitre 2 – Article

Evidences that low-level visual information is stored specifically with associated higher-level features during recognition memory

Larouche, J-M[1], Gosselin, F.[1]
1. Département de psychologie, Université de Montréal

Introduction

The human capacity to recognize complex visual objects depends on a hierarchical sequence of brain areas known as the ventral stream. Recent studies have demonstrated that some specific visual features in the lower level areas are also represented in the higher levels of this visual stream (Bones, Ahmad & Buchsbaum, 2020; Hong, Yamins, Majaj & DiCarlo, 2016) and even observed that a similar mechanism emerges in Deep convolutional neural networks (CNNs) trained for object classification, where higher level layers represent not just the object categories they are trained to classify but also lower level features such as the shape and the position of objects (Hong, Yamins, Majaj & DiCarlo, 2016; Cichy & Kaiser, 2019; Rajalingham, Rishi, et al., 2018). These discoveries suggest a top-down directionality in the human ventral stream, just like we can observe in CNNs architectures trained for object recognition. Then, should we observe the same directionality during the activation of a memory trace for more complex lower level representations? In this article, we observe that combinations of low-level features (spatial-frequencies and orientations) are encoded specifically with the higher level representations they were associated with during learning of higher level information (like the

facial features in our experiment) and these low-level representations cannot be retrieved without the learned higher level visual representation available and reactivated.

Multiple studies interested in the directionality of the human memory encoding model compared the influence of low and high level features on recognition interferences using various experimental protocols and arrived to a similar conclusion: unlike high-level information like object categories, low-level representation doesn't impact recognition memory interferences (Baddeley, 1966b; Ishai et Sagi, 1995, 1997; Konkle et al. 2010a; Konkle et al. 2010b). Surprisingly according to the researchers, interferences during a recognition task were not predicted by their perceptual similarity measures. The interferences were similar in the object categories with perceptually distinct objects or not. Therefore, they concluded that category information is a critical part of the long-term memory encoding model, suggesting that distinctions along these high-level categorical dimensions provide "conceptual hooks" allowing to retrieve the full mnemonic traces, including the perceptual information.

However, we contend that all these studies on memory interference were ill suited to assess how low-level properties are encoded in long-term memory. Indeed, in all these experiments on memory interference the stimuli were learned over only one trial or few trials, but learning perceptual information in long-term memory is often a slow process that is statistically driven (Karni and Sagi, 1993). These low-level brain regions − allowing, for example, visual perception (V1-V2) − demonstrate a form of neuronal plasticity resulting from informational feedback directed either by higher-level representations in the hierarchy or by rewards from external regions routed among others by the striatum which − unlike the medial temporal lobe − receives and sends information almost everywhere in the brain (McGeorge and Faull 1989, Petrov, Dosher

et al. 2005). In fact, several iterations of exposure to the stimuli are necessary using prediction errors in order to correctly encode the perceptual information useful for a requested task. Moreover, according to the reverse hierarchy, learning perceptual information is theorized as a top-down guided process as well, which begins in higher-level areas of the visual systems and when it does not suffice for learning, progresses backwards to the input level to access a better signal-to-noise ratio (Ahissar and Hochtein, 2002, 2004). In this article, we attempted to create an association between low-level information and a behavioral response. If the information is useful for the task and learned over multiple iterations, we expect that, in this situation, the similarity of low-level information (spatial frequencies and orientations composing the facial features in our study) will predict recognition interference between learned objects (first hypothesis).

Although, the low-level information should not impact recognition interference independently from the high-level features associated with the low-level information during the learning phase − recalling low-level information being dependent on the associated higher-level representation. In other words, low-level perceptual similarity associated with new high-level features (facial features not learned during the experiment) should not impact memory interference and this, even if perceived during the same recognition task and similar in terms of high-level features (second hypothesis). If the low-level similarity is predicting memory interferences independently of its association with higher level features during learning, it would infirm the reverse hierarchy theory. Otherwise, we will be able to consolidate the theory.

To test our hypotheses, we used three complementary filters (green, red and blue in Figure 1) applied in the stimuli' frequency domain to sample combinations of different spatial frequencies with different orientations. These stimuli produced very distinct activations in the

striate cortex and other early visual visual areas (De Valois et De Valois, 1990) but similar activations in high-level visual areas. In other words, by filtering lists of images with different filters, we created quasi-metamers, i.e. lists of stimuli that are evenly similar in higher level, but different between the lists in terms of lower-level features.

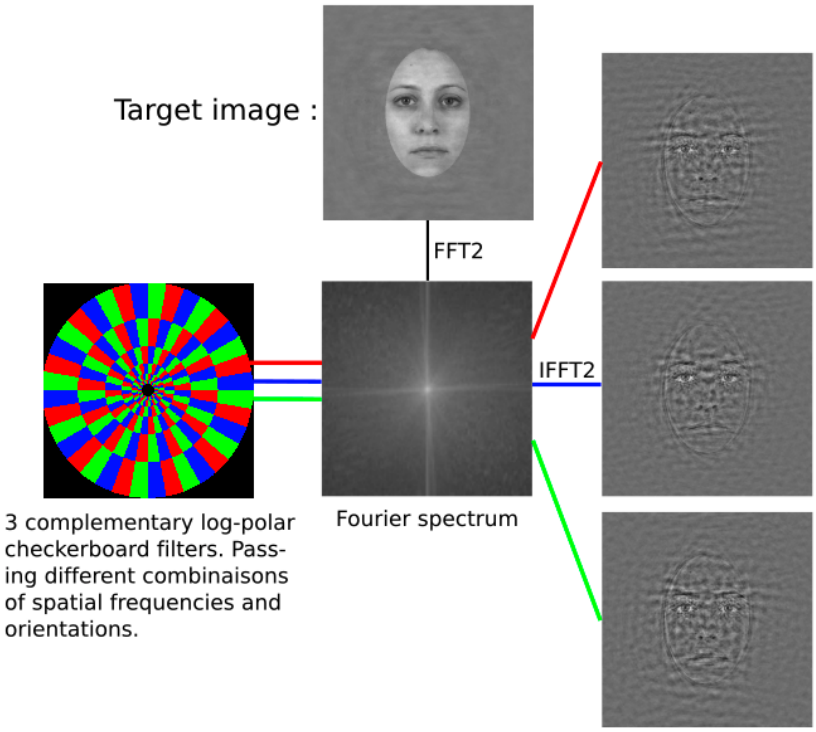

Figure 1. − Three complementary log-polar checkerboard filters applied in the Fourier domain

Caption: Three complementary filters (green, red and blue) when applied on the Fourier spectrum of an image sample three different combinations of spatial frequencies and orientations (36 different combinations of 6 SFs at 6 orientations for each filter). Images filtered with the same filter in the Fourier domain are more similar in V1 than images filtered with different filters. Also, two faces filtered with the same filter are not more alike than two faces filtered with different filters (resulting in quasi-metamers).

Two groups learned two sets of faces − composed either of the same or of two different combinations of SF and orientations (same or different filters between the two sets). On day one, participants learned the stimuli of the first set in a target/non-target task, then the stimuli in the second set. Leaning tasks consisted of target/non-target classification with auditory feedback during multiple iterations − all the non-target faces were filtered with a third filter, making the low-level information useful for the task. On day two, in a task with three alternatives, the participants distinguished the faces from list A, list B and a new non-target set (filtered with the three low-level complementary filters randomly chosen). Interferences between list A, list B and new-target faces were measured for every participant on day 2.

Hypotheses

1. The low-level visual properties of the learned faces should impact long-term memory interference when the useful information is learned over multiple iterations: there should be evidence that there is more interference between the first and second sets in the group with similar low-level properties in the two learning sets than in the group with different low-level properties in the two learning sets on day 2.

2. The low-level visual properties of the new non-target faces should not impact memory interference because the human memory encoding model uses higher-level features to reactivate the low-level information: there should be evidence that there is no difference between the interferences of the non-target face with similar or different low-level properties from the two face sets on day 2.

Results

In this experiment, there are three types of interference. First, the proactive interference (PI) consists in classifying a face belonging to the first set learned by a participant as belonging to the second set learned by the participant. Second, the retroactive interference (RI), on the contrary, consists in classifying a face from the second set learned by a participant as belonging to the first set learned by the participant. Third, and finally, the non-target interference (NTI) consists in classifying one of the 24 new non-target faces used during the test as belonging to either face list A, or B. We calculated the average number of interferences for each stimulus and each participant. PIs and RIs are used to test our first hypothesis (that low-level similarities impact memory interferences), and NTIs are used for our second hypothesis (that low-level similarities should not impact memory interferences for new non-target faces, with new high-level facial features). We used Bayesian factors to quantify both the evidence for our experimental hypothesis on the impact of low-level properties on long-term memory interferences (BF10) and the evidence for the null for our second hypothesis on the absence of an impact from low-level similarity when not seen with the correct higher-level features learned during the experiment (BF01).

Participants with accuracy (during the three alternative choices task) that were not significantly different from chances (using bayesian t-test) were excluded from the analysis (9 outlier participants out of 93; 66 womens). The average interference rate for an image was 0.2563 (std: 0.11) for the first group and 0.21 (std: 0.12) for the second. The average non-target

interference is 0.4299 (std: 0.07) for the same filter condition and 0.7268 (std: 0.03) for the different filter condition.

We used Bayesian factors to quantify both the evidence for our experimental hypothesis on the impact of low-level properties on long-term memory interferences (BF10) and the evidence for the null for our second hypothesis on the absence of an impact from low-level similarity when not seen with the correct higher-level features learned during the experiment (BF01). The number of interferences (proactive and retroactive interferences combined) between the two learnings is decisively greater for the identical low-level filter condition than for the different low-level filter condition (BF10: 4.1355; p=0.0073). In fact, the Bayes factor indicates moderate evidences in favor of our experimental hypothesis, namely that memory interference increases as a function of the low-level visual properties' similarity (SFs and orientations) despite little high-level perceptual differences (quasi-metamer) − when the learning involves several iterations and the information is useful for the task.

For the 24 new non-target faces during the recall task on day two, there is evidence that the interferences are the same for all non-target faces whether they were filtered like or unlike the target faces (BF10: 0; BF01: Inf; p=1). The new non-target faces were previously unseen faces randomly filtered with one of our three complementary log-polar checkerboard filters. Even when they were more similar at low-level (depending on the filter), they were all dissimilar from the learned target faces at high-level (having different facial features from those trained to remember). These results suggest that the learning of early visual representations needs to be contextualized by a higher-level representation (any perceptual features leading to the recognition of the face) to be remembered and affect memory interferences. Then, as stipulated by our

hypotheses, SFs and orientations similarity in early visual cortex impact long-term memory interferences (first hypothesis), but only when these low-level firing patterns are associated with the correct higher-level visual features learned during the experiment (second hypothesis).

Discussion

Our first hypothesis revisited the effect of low-level visual properties similarity (SFs and orientations; in early visual cortex) on long-term memory interferences with a protocol that is more suitable for promoting perceptual learning. By using three complementary log-polar checkerboard filters sampling different combinations of SFs and orientations − quasi-metamers −, we created two experimental groups: one where learning lists A and B had similar low-level properties (identical filters), and the other where learning lists A and B had dissimilar low-level properties (different filters). Bayesian factors demonstrated decisive evidence in favor of our first experimental hypothesis, that interference increases according to the similarity of low-level processing contrary to what has been reported by other studies (Baddeley, 1966a, 1966b; Ishai and Sagi, 1995, 1997a, 1997b; Konkle et al., 2010b).

Our second hypothesis stipulated that remembering low-level properties should be specific to the higher-level representations associated during learning. Indeed, our experiment indicates strong evidences that low-level similarity of the new non-target faces did not impact the retrieval of the learned faces from the set A or B. Also, our findings are consistent with the reverse hierarchy theory (RHT) and many other studies on higher-level influences on processing and memory (Ahissar and Hochstein, 2002, 2004; Li et al., 2004; Li et al., 2008; McManus et al., 2011). The scientific literature considers learning as a top-down process, where influences of

attention, expectations, perceptual tasks, etc. drive the contextualization and adaptation of the neurons processing functions (Gilbert et Li, 2013; Li et al., 2004). Evidence in this study demonstrates that perceptual learning is indeed a top-down process, prioritizing the higher-level representation to structure the retrieval of long-term memories. To access the correct low-level representation, the correct higher-level conceptual key − associated with the low-level information during learning − is required. This is well explained by the fact that perceptual information about our world is vast and complex, and specific low-level representations are needed in different contexts. Representing information in a hierarchy allows us to think and generalize to different situations. The best way to recover the correct low-level information seems to be by using simpler and more general high-level concepts.

In order to ensure that the reported effect is not explained by slight residual higher-level differences associated with the three complementary filters, we ran a simulation using a deep neural network pre-trained on millions of images (AlexNet; Krizhevsky and Hinton, 2012). By removing the share of interference explained by the similarity of the target images at the different levels of the model, we observed that the intergroup difference is mostly explained by the similarity in the second layer of the model − reported being most correlated with V1 activity (Cichy et al., 2016). Also, the fact that non-target faces filtered differently or not from target faces did not impact memory interference is an indication that our filters are specific to low-level processing and impact minimally higher-level similarity. Thus, we conclude that similarity of low-level visual properties (~V1) influences interference in long-term perceptual memory when

the useful information for the task is learned over several iterations and associated with the correct higher-level features during recall.

Beyond this crude hierarchical distinction, our experiment does not allow inferring precise processing loci in the human brain. For example, the striatum may be involved in our effect because there is a known link between it and V1, but our effect may also be present in V1-V2 which is directly linked to the perirhinal cortex in the medial temporal lobe (Peterson et al., 2012; Clavagnier et al., 2014). Knowing that the medial temporal lobe has a demonstrated important role in statistical learning (McClelland et al., 1995; Shapiro et al., 2016), the two systems are possible candidates, which probably works in interaction for our effect. In the future, more precise metameric stimuli and fMRI decoding will be needed to understand how the visual properties of objects are encoded by the different learning systems throughout the entire visual hierarchy.

That being said, now we know there is evidences − according to our Bayesian factors and our quasi-metamers − that low-level visual properties' similarity (SFs and orientations; ~V1) increase the number of interferences in long-term memory when the unconscious information is useful for the task and learn over several iterations. More importantly, our experiment provides psychophysical evidence that low-level information is encoded specifically with the higher-level features associated with the low-level information during learning. These high-level features are necessary to retrieve the specific low-levels information, even during recognition memory.

A recent study by Bone et al. (2020) used a new FMRI decoding approach (FSIC) and demonstrated the presence of low-level representations within the higher-order areas in the ventral and dorsal stream, as well as the frontal cortex. Like the nodes of a feed-forward CNN

performing visual classification and localization tasks, the brain neurons' receptive fields are organized in such a way that lower-level layers have smaller receptive fields and weak semantics; and that higher-level layers have larger receptive fields with strong semantic value. In result, there is a loss of fine details essential for some tasks. They posited that the presence of these representations in higher level facilitate object classification, attentional allocation and motor planning in tasks that require both accurate semantic and fine perceptual details during episodic memory. Our results are consistent with Bone et al. (2020) discovery. Other fMRI studies could observe if these higher level representations of lower level features are the conceptual keys to retrieve lower level information in the early visual cortex or specific representation of lower level information.

In conclusion, our findings explain apparent contradictions in the litterature, demonstrate an evident directionality in the human memory encoding model and help consolidate the relation between the human mind and deep neural networks.

Method

Participants

Ninety-three neurotypical adults (66 females) between the age of 18 and 30 years old (mean: 22.40; std: 3.6), with normal or corrected to normal vision, participated in our study. All participants signed a consent form approved by the Université de Montréal ethics committee and received 30 $ as monetary compensation. Participants were recruited with Facebook posts on research participation groups. Participants were randomly assigned to one of two subject groups: the first group had to learn two sets of face stimuli with similar low-level properties (N = 48; 37

females) while the second one had to learn two sets of face stimuli with dissimilar low-level properties (N = 45; 35 females). We excluded 9 participants from the analyses because their mean accuracy was significantly below the statistical threshold for chance during the three alternative choices recall task. The learning tasks were very difficult by design to promote the learning of low-level representations.

Stimuli

Stimuli were created by filtering faces from the Chicago Face Database (Ma et al. 2015). Faces with attributes that facilitate recognition (mole, visible hair, etc.) were eliminated. An elliptical mask was applied to the faces to show only their internal features. SFs content was equated across face images with SHINE Toolbox (Willenbockel et al. 2010).

The idea of this study was to create two sets of faces (set A and set B; in the same category), but with similar or dissimilar low-level properties between the two learning sets. The visual modules of the striate cortex represent information specifically for certain SFs at certain orientations (De Valois et De Valois, 1990), but different sinewave gratings can lead to similar categorizations at higher-level (quasi-metamers). The 2D fast Fourier transformation (FFT2) of an image allows its representation in the frequency domain − where for each orientation, the lowest frequencies are in the center and the highest are in the periphery on the spectrum of magnitudes. Our three complementary filters (green, red and blue in Figure 1) sample 36 different combinations of 6 spatial frequencies at 6 orientations for each filter. All filter cutoff frequency under 13 cycle/image − to prevent aliasing at the center− and above 250 cycle/image (images size being 500 by 500 pixels; for equivalent sampling in all orientations). One should

expect that images filtered with the same filter in the Fourier domain are more similar in V1 than images filtered with different filters. A face filtered with the three filters produces images that are slightly different to the naked eye but not nearly as much as another face filtered with these filters. We have shown that something analogous happens in AlexNet (see Simulation section). In other words, a face filtered with the three filters results in quasi-metamers.

Procedure

All participants completed three tasks over a two-day period: on day 1, subjects learned the list of faces A (task 1) and the list of faces B (task 2) using a target/non-target task; on day 2 (at least 24 hours later), the participants identified the faces from list A, list B and new non-target faces in a three-alternative task (task 3), making it possible to measure proactive (PI), retroactive (RI) and non-target interference (NTI) at the same time. Twenty-four hours are insufficient to consolidate memory completely (Squire, 1986). However, this delay still ensures the presence of a trace in long-term memory (Karni and Sagi, 1993). Proactive interference (PI) refers to the classification of an image from the first learning as belonging to the second, and retroactive interference (RI) refers to the classification of an image from the second learning as belonging to the first. We also measured the interference from the 24 new non-target faces classified as a learned target (non-target interference; NTI). As mentioned, the learning tasks consisted of target / non-target classifications.

Initially, the 12 targets of the task were presented to participants for 8 seconds with an interval of 2 seconds between each image. Then, an image appeared in the middle of the screen

for 2 seconds. After the image disappeared, the participant had to judge whether the face just shown was a target or not by pressing the appropriate keyboard key. Different feedback and delays indicated to participants whether the response was correct (high tone; 0.5 second) or incorrect (low tone; 0.75 second). For the two learning tasks, each of the 12 images returned 30 times and 24 non-target images returned 15 times. The recall task on day 2 consisted of a classification task with three alternatives, but without feedback (delay was always equal to 0.5 second). In this task, the 12 targets images A and B and 24 new non-target faces were shown 20 times each. The participant had to discriminate target A, target B and non-target faces − classification errors of target images constitute interference in long-term memory. The experimental group were determined randomly for each participant. For every task, the non-target images and the order of presentation of the images were also determined randomly.

The 12 targets A and target B faces were the same for all participants but filtered differently. To minimize the possibility of affecting high-level processing with the low-level sampling, the target faces chosen in the experiment are those whose three filtered images are the most correlated with each other, as well as with the three filtering of an average image (representing all faces in the experiment). Knowing that perceptual learning only encodes information useful for the task (Ahissar and Hochtein, 2002, 2004), filtering non-target images with the third filter (during learnings) made our low-level property sampling useful for solving our task. The different filters used were determined randomly and each face was filtered with a unique filter throughout the entire experiment for each participant.

During the recall task on day 2, the 24 new non-target faces were filtered with the three different filters (8 random images for each filter) and the same non-target images did not come

back through tasks. This allowed us to measure if the learning of low-level properties is independent or specific to the higher-level representation of the images learned on day 1. On the one hand, if non-target images interfere more with target images of the same filter then of different filter, this means that the influence of low-level properties on behavior is independent of the higher-level representation of the target faces. If there is no difference between the filtering conditions, it means that low-level properties are specific to higher-level representation, which contextualize the retrieval of the low-level representation.

Apparatus

The experimental programs ran on Mac Pro (Apple Inc.) computers in the Matlab (Mathworks Inc.) environment, using functions from the Psychophysics Toolbox (Brainard, 1997). All stimuli were presented on Asus VG278H monitors (1920 × 1080 pixels at 120 Hz), calibrated to allow a linear manipulation of luminance. Luminance ranged from 1.6 cd/m^2 to 159 cd/m^2. During the experiment, participants were seated in a room painted in black dimly lit only by light from the computer monitor. The image spanned 16 cm (12 deg of visual angle) on the computer monitors. Viewing distance was maintained at 76.11 cm with a chin-rest. To do the necessary computations, we used the bayesFactor toolbox (https://klabhub.github.io/bayesFactor/) implemented in Matlab and based on Kass and Raftery (1995) mathematical framework.

Simulation

In the visual hierarchical stream, low-level information influences the composition of invariant representations in higher level areas. Even if our filters are not consciously distinguishable from each other, there is probably some of the effect that is due to nonlinear combinations of low-level visual properties in the higher-level processing. To test whether the reported effect is consistent with low-level processing (SFs and orientations), we used the analog properties between the human ventral stream and a pre-trained deep learning network (Alexnet; see the annexe for a description of the model; Krizhevsky and Hinton, 2012). This analogy does not allow inferring precise loci in the human brain that fit with our behavioral data. However, it is a rough approximation for the invariance of visual processing, that is, how low-level properties influence the invariant composition of facial features.

From this, we extracted the activations for the 24 target faces, filtered by the three complementary filters (72 images), in the eight layers of the model. Then, for all the filtering conditions (9; having three possible filters for the two learning sets), we created a dissimilarity vector (1-Pearson correlation) between the activation of an image and of other images in the same and the opposite learning sets. Thereafter, we tried to predict the average number of interferences for all participant images in function of these vectors, and this, independently for all layers of Alexnet (Krizhevsky and Hinton, 2012). Several predictive models have been tested (Neural Network, Support Vector Regression, Gaussian Process Regression; etc.). However, multiple linear regression model was retained, having a better fit in general (R2; Layer 1: 0.2180, Layer 2: 0.3425, Layer 3: 0.2768, Layer 4: 0.2284, Layer 5: 0.2522, Layer 6 : 0.2191, Layer 7: 0.2045,

Layer 8: 0.1354), in addition to being the most parsimonious solution. In order to explore which level of representational distance further explains the intergroup difference identified above, we removed the interference share predicted by these eight linear models (one for every Alexnet level) for all participants. Subsequently, by comparing our two groups again for each layer subtraction, we can observe the accumulation of evidences that there is no longer any intergroup differences (BF01) − after removing the interference share explained by the dissimilarity at each level of processing (see Figure 5 a).

There is evidence that the effect of our first hypothesis is explained by low-level processing similarity. Indeed, the target faces similarity in the second layer of Alexnet mostly explains the intergroup difference in our first hypothesis. In a study conducted by (Cichy et al., 2016), they demonstrated that the second layer of an Object DNN like Alexnet is most correlated with activations in V1. This correlation explains the idea that similarity in the second layer most significantly predicts the difference between the two groups in figure 5 a. Thus, our confirmatory simulation leads to the idea that the intergroup difference is mostly explained by the similarity of the low-level processing (in layer 2) − which is in agreement with our experimental hypothesis, that low-level similarity in the hierarchical visual cortex can impact long-term memory interferences.

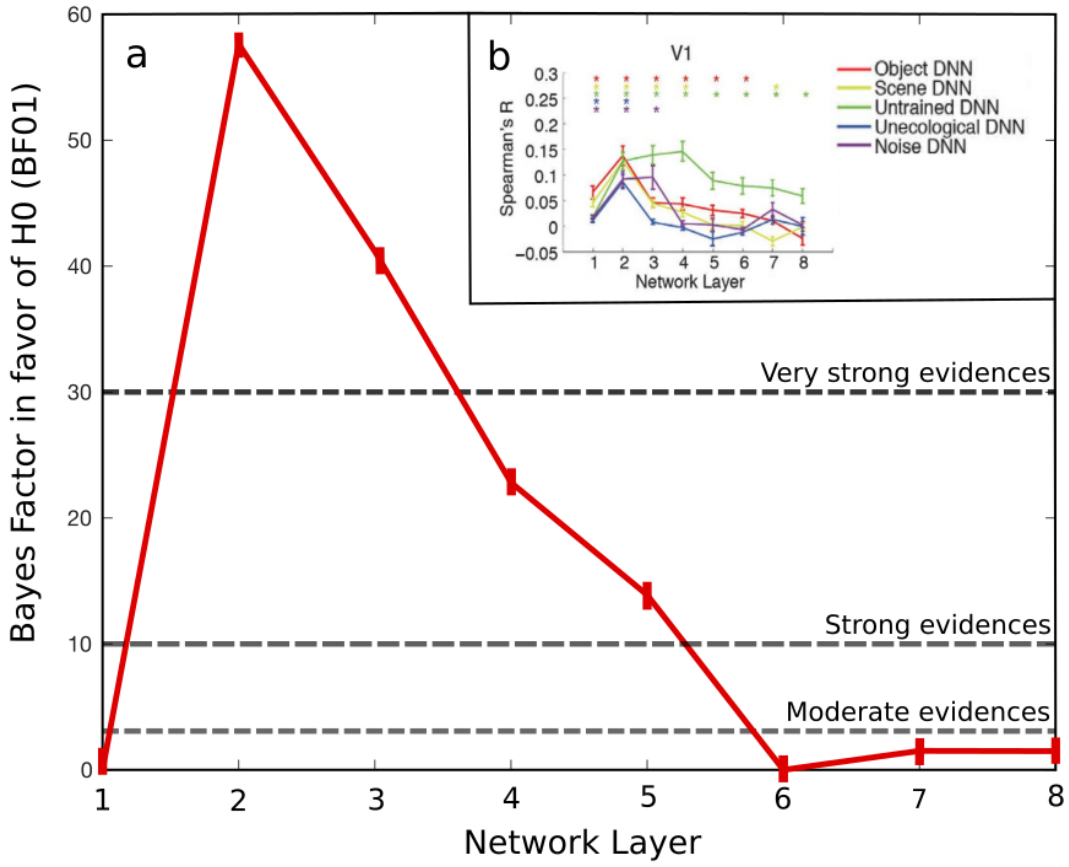

Figure 2. − Evidences in favor of the null hypothesis after removing the interference share predicted by each layer similarity

Caption: In the figure 5, **a)** represents the evidences that there are no differences between the two experimental groups (Bf01) after removing for each level of Alexnet (Krizhevsky and Hinton, 2012) the participant interference shares predicted by the dissimilarity between an face and all other target faces. While **b)** is a graph borrowed from (Cichy et al., 2016) showing that the activity in V1 is generally more correlated in their experiment with the activity of the second layer of object DNN like Alexnet (in red).

Chapitre 3 – Discussion générale

Notre première hypothèse a revisité l'effet de la similarité des propriétés visuelles de bas niveau (FS et orientations ; dans le cortex visuel précoce) sur les interférences de la mémoire à long terme avec un protocole plus adapté pour favoriser l'apprentissage perceptif. En utilisant trois filtres en damier log-polaires complémentaires échantillonnant différentes combinaisons de fréquences spatiales et d'orientations − quasi-métamères −, nous avons créé deux groupes expérimentaux : l'un où les listes d'e visages A et B avaient des propriétés de bas niveau similaires (filtres identiques), et l'autre où les listes d'apprentissage A et B avaient des propriétés de bas niveau différentes (filtres différents). Les facteurs bayésiens ont démontré des preuves en faveur de notre première hypothèse expérimentale, à savoir que l'interférence augmente en fonction de la similarité du traitement de bas niveau contrairement à ce qui a été rapporté par d'autres études (Baddeley, 1966a, 1966b ; Ishai et Sagi, 1995, 1997a, 1997b ; Konkle et al., 2010b).

Notre hypothèse secondaire stipulait que la mémorisation des propriétés de bas niveau devait être spécifique aux représentations de plus haut niveau associées lors de l'apprentissage. En effet, notre expérience indique des preuves que la similarité de bas niveau des nouveaux visages non cibles n'a pas eu d'impact sur la récupération des visages appris à partir de l'ensemble A ou B. Ces résultats sont cohérents avec la théorie de la hiérarchie inverse (RHT) et de nombreuses autres études sur les influences de haut niveau sur le traitement et la mémoire

(Ahissar et Hochstein, 2002, 2004 ; Li et al., 2004 ; Li et al., 2008 ; McManus et al., 2011). La littérature scientifique considère l'apprentissage comme un processus descendant, où les influences de l'attention, des attentes, des tâches perceptives, etc. conduisent la contextualisation et l'adaptation des fonctions de traitement des neurones (Gilbert et Li, 2013 ; Li et al., 2004). Cette étude démontre que l'apprentissage perceptif est en effet un processus descendant, donnant la priorité à la représentation de niveau supérieur pour structurer la récupération des souvenirs à long terme. Pour accéder à la représentation de bas niveau correcte, la bonne clé conceptuelle de niveau supérieur - associée aux informations de bas niveau lors de l'apprentissage - est requise. Cela s'explique bien par le fait que les informations perceptuelles sur notre monde sont vastes et complexes, et que des représentations spécifiques de bas niveau sont nécessaires dans différents contextes. Représenter les informations dans une hiérarchie nous permet de penser et de généraliser à différentes situations. La meilleure façon de récupérer les informations de bas niveau correctes semble être d'utiliser des concepts de haut niveau plus simples et plus généraux.

Afin de nous assurer que l'effet rapporté ci-haut n'est pas expliqué par des différences de similarités résiduelles de haut niveau associées aux trois filtres complémentaires, nous avons effectué une simulation en utilisant un réseau de neurones d'apprentissage profond pré entraîné sur des millions d'images (AlexNet; Krizhevsky and Hinton, 2012). En supprimant la part d'interférence expliquée par la similarité des images cibles aux différents niveaux du modèle, nous avons observé que la différence intergroupe s'explique principalement par la similarité dans la deuxième couche du modèle − rapportée comme étant la plus corrélée avec l'activité de V1 (Cichy et al., 2016). De plus, le fait que les visages non-cibles filtrées différemment ou non des visages cibles n'aient pas d'impact sur les interférences en mémoire est une indication importante

que nos filtres sont en fait spécifiques au traitement de bas niveau et ont un impact minimal sur la similarité des propriétés faciales de plus haut niveau. Nous concluons ainsi que la similarité des propriétés visuelles de bas niveau (~V1) influence les interférences en mémoire à long terme lorsque ces informations utiles pour la tâche sont apprises sur plusieurs itérations et associées aux bonnes caractéristiques de haut-niveau lors du rappel − celles ayant été associées à l'information de bas niveau durant l'expérience.

Au-delà de cette distinction hiérarchique (bas et haut niveau), notre expérience ne permet pas de déduire des lieux de traitement précis dans le cerveau humain. Par exemple, le striatum peut être impliqué dans notre effet car il existe un lien connu entre lui et V1, mais notre effet peut également être présent dans V1-V2 qui est directement lié au cortex périrhinal dans le lobe temporal médial (Peterson et al ., 2012 ; Clavagnier et al., 2014). Sachant que le lobe temporal médian a un rôle important démontré dans l'apprentissage statistique (McClelland et al., 1995 ; Shapiro et al., 2016), les deux systèmes sont des candidats possibles, ce qui fonctionne probablement en interaction pour notre effet. À l'avenir, des stimuli métamériques plus précis et un décodage IRMf seront nécessaires pour comprendre comment les propriétés visuelles des objets sont codées par les différents systèmes d'apprentissage dans toute la hiérarchie visuelle de la voie ventrale.

Cela étant dit, nous savons maintenant qu'il existe des preuves - selon nos facteurs bayésiens et nos quasi-métamères - que la similarité des propriétés visuelles de bas niveau (FS et orientations ; ~V1) augmente le nombre d'interférences dans la mémoire à long terme lorsque le l'information inconsciente est utile pour la tâche et apprend sur plusieurs itérations. Plus important encore, notre expérience fournit des preuves psychophysiques que les informations de

bas niveau sont codées spécifiquement avec les caractéristiques de haut niveau associées aux informations de bas niveau pendant l'apprentissage. Ces fonctionnalités de haut niveau sont nécessaires pour récupérer les informations spécifiques de bas niveau, même pendant la mémoire de reconnaissance.

Une étude récente de Bone et al. (2020) a utilisé une nouvelle approche de décodage IRMf (FSIC) et a démontré la présence de représentations de bas niveau dans les zones d'ordre supérieur des flux ventral et dorsal, ainsi que du cortex frontal. Ils ont postulé que la présence de ces représentations à un niveau supérieur facilite la classification des objets, l'allocation attentionnelle et la planification motrice dans les tâches qui nécessitent à la fois des détails sémantiques précis et des détails perceptifs fins pendant la mémoire épisodique. Nos résultats sont en accord avec les résultats de Bone et al. (2020). D'autres études d'IRMf pourraient maintenant observer si ces représentations de caractéristiques perceptuelles dans les niveaux supérieures de la voie ventrale sont en fait des clés conceptuelles permettant de récupérer des informations de niveau inférieur dans le cortex visuel précoce ou une représentation spécifique d'informations de niveau inférieur.

En conclusion, nos résultats expliquent les contradictions apparentes dans la littérature, démontrent une directionnalité évidente dans le modèle d'encodage de la mémoire humaine et aident à consolider la relation entre l'esprit humain et les réseaux de neurones profonds. Selon Borges, dans le second livre du *Naturalis Historia*, Pliny écrit sur la mémoire en la décrivant comme « un avantage nécessaire à la vie ». Maintenant, cette nécessité pour la vie ne repose point sur l’idée d’une mémoire infaillible comme celle de Funes − permettant une représentation parfaite du monde extérieur; celle-ci repose en fait sur une mémoire moins exacte et plus

générale, la nôtre. Intuitivement, on explique cette mémoire imparfaite par l'idée d'un monde complexe, de notre capacité limitée. Toutefois, structurer la mémoire à long terme à partir de nos représentations plus abstraites et générales du monde extérieur − encodant la régularité des événements spatio-temporels de notre environnement − nous permet de par la suite généraliser nos représentations à des situations futures. La vie étant dans l'obligation constante de s'adapter à un monde en changement, sans quoi elle ne peut plus être, structurer l'encodage et le rappel de l'information en mémoire à long terme en utilisant le plus général pour retrouver le particulier représente un avantage évident afin d'adapter les différentes représentations d'un système vivant à son environnement futur. En réalité, c'est même là l'avantage nécessaire que peut représenter la mémoire pour la vie.

Références bibliographiques

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, *8*(10), 457-464.

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, *8*(10), 457-464.

Anderson, M. C., Green, C., & McCulloch, K. C. (2000). Similarity and inhibition in long-term memory: evidence for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1141.

Baddeley, A. D. (1966a). The influence of acoustic and semantic similarity on long-term memory for word sequences. *The Quarterly journal of experimental psychology*, *18*(4), 302-309.

Baddeley, A. D. (1966b). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly journal of experimental psychology*, *18*(4), 362-365.

Bilodeau, I. M., & Schlosberg, H. (1951). Similarity in stimulating conditions as a variable in retroactive inhibition. *Journal of experimental psychology*, *41*(3), 199.

Borges, J. L. (1957). Funes ou la mémoire. *Fictions*, 109-118.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of vision*, *11*(5), 4-4.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433-436.

Clavagnier, S., Falchier, A., & Kennedy, H. (2004). Long-distance feedback projections to area V1: implications for multisensory integration, spatial awareness, and visual consciousness. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 117-126.

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.

Crowder, R. G. (2014). *Principles of learning and memory: Classic edition*. Psychology Press.

DeValois, R. L., & DeValois, K. K. (1990). *Spatial vision* (Vol. 14). Oxford university press.

Dosher, B. A., & Lu, Z. L. (1999). Mechanisms of perceptual learning. *Vision research*, *39*(19), 3197-3221.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350-363.

Humphreys, G. W., & Bruce, V. (1989). *Visual cognition: Computational, experimental and neuropsychological perspectives*. Psychology Press.

Ishai, A., & Sagi, D. (1995). Common mechanisms of visual imagery and perception. *Science*, *268*(5218), 1772-1774

Ishai, A., & Sagi, D. (1997a). Visual imagery facilitates visual perception: Psychophysical evidence. *Journal of Cognitive Neuroscience*, *9*(4), 476-489.

Ishai, A., & Sagi, D. (1997b). Visual imagery: Effects of short-and long-term memory. *Journal of Cognitive Neuroscience*, *9*(6), 734-742.

Karni, A., & Sagi, D. (1993). The time course of learning a visual skill. *Nature*, *365*(6443), 250-252.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773-795.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Kosslyn, S. M. (1980). *Image and mind*. Harvard University Press

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010a). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological science*, *21*(11), 1551-1556.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010b). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Law, C. T., & Gold, J. I. (2009). Reinforcement learning can account for associative and

perceptual learning on a visual-decision task. *Nature neuroscience*, *12*(5), 655-663.

Li, W., Piëch, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience*, *7*(6), 651-657.

Li, W., Piëch, V., & Gilbert, C. D. (2008). Learning to link visual contours. *Neuron*, *57*(3), 442-451.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, *47*(4), 1122-1135.

Magnussen, S., & Greenlee, M. W. (1999). The psychophysics of perceptual memory. *Psychological research*, *62*(2-3), 81-92.

Mahut, H., Zola-Morgan, S. T. U. A. R. T., & Moss, M. (1982). Hippocampal resections impair associative learning and recognition memory in the monkey. *Journal of Neuroscience*, *2*(9), 1214-1220.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419.

McGaugh, J. L. (2000). Memory--a century of consolidation. *Science*, *287*(5451), 248-251.

McGeorge, A. J., & Faull, R. L. M. (1989). The organization of the projection from the cerebral cortex to the striatum in the rat. *Neuroscience*, *29*(3), 503-537.

McManus, J. N., Li, W., & Gilbert, C. D. (2011). Adaptive shape processing in primary visual cortex. *Proceedings of the National Academy of Sciences*, *108*(24), 9739-9746.

Melton, A. W., & Von Lackum, W. J. (1941). Retroactive and proactive inhibition in retention: Evidence for a two-factor theory of retroactive inhibition. *The American Journal of Psychology*, *54*(2), 157-173.

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760-767.

Müller, G. E., & Pilzecker, A. (1900). *Experimentelle beiträge zur lehre vom gedächtniss* (Vol. 1). JA Barth.

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature neuroscience*, *5*(11), 1203-1209.

Osgood, C. E. (1946). Meaningful similarity and interference in learning. *Journal of Experimental Psychology*, *36*(4), 277.

Packard, M. G., & Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual review of neuroscience*, *25*(1), 563-593.

Peterson, M. A., Cacciamani, L., Barense, M. D., & Scalf, P. E. (2012). The perirhinal cortex modulates V2 activity in response to the agreement between part familiarity and configuration familiarity. *Hippocampus*, *22*(10), 1965-1977.

Petrov, A. A., Dosher, B. A., & Lu, Z. L. (2005). The dynamics of perceptual learning: an incremental reweighting model. *Psychological review*, *112*(4), 715.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252.

Schapiro, A. C., Turk‐Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3-8.

Simon, H. A. (1991). The architecture of complexity. In *Facets of systems science* (pp. 457-476). Springer, Boston, MA.

Squire, L. R. (1986). Mechanisms of memory. *Science*, *232*(4758), 1612-1619.

Wen, H., Shi, J., Zhang, Y., Lu, K. H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, *28*(12), 4136-4160.

Wimmer, G. E., Braun, E. K., Daw, N. D., & Shohamy, D. (2014). Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *Journal of Neuroscience*, *34*(45), 14901-14912.

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behavior research methods*, *42*(3), 671-684.

Annexes

We used Alexnet for two main reasons, firstly, this model is relatively simple − compared to more recent models − and secondly, its relationship with the human visual system is very well documented (Khaligh-Razavi, 2014; Cichy et al., 2016; Wen et al., 2018). The model was pre-trained on 1.3 million natural images (ImageNet; Krizhevsky and Hinton, 2012; Russakovsky et al., 2015) allowing the classification of 1000 different categories and almost reaching human performance. Alexnet consists of 8 layers stacked in a hierarchical architecture, where the previous layers transmit information to the next layer (Krizhevsky and Hinton, 2012). The first five layers are convolutions, while the last three are fully connected layers. While the fully connected layers (fc6, fc7 and fc8) are made up of vector (sizes of 4096, 4096 and 1000 units respectively), the convolutional layers have dimensions of: layer 1 (conv1) 96 × 55 × 55 (96 features, more than 55 × 55 retinotopic units), layer 2 (conv2) 256 × 27 × 27, layer 3 (conv3) 384 × 13 × 13, layer 4 (conv4) 384 × 13 × 13 and layer 5 (conv5) 256 × 13 × 13 (see figure 3).

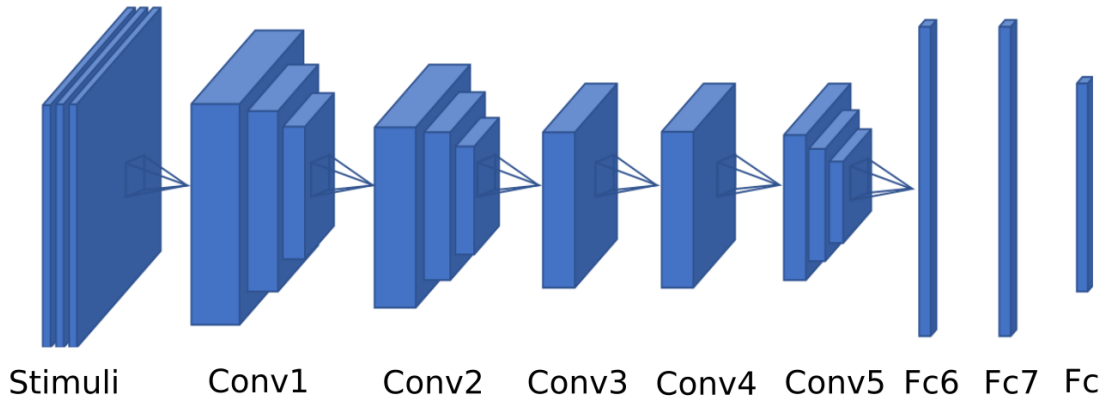

Figure 3. − Alexnet hierarchical architecture

Caption: This figure represents the hierarchical architecture of Alexnet model, which have 8 layers, 5 convolutions and 3 fully connected layers.