THESE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *L'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Neurosciences*

Présentée et soutenue par *Sébastien Crouzet*
Le *12 juillet 2010*

Titre : *Jeter un regard sur une phase précoce des traitements visuels*

JURY

Françoise VITU-THIBAULT - CNRS, Université de Provence, Marseille	*Rapporteur*
Guillaume ROUSSELET - University of Glasgow	*Rapporteur*
Olivier PASCALIS - CNRS, Université Pierre Mendès France, Grenoble	*Examinateur*
Bruno ROSSION - Université Catholique de Louvain	*Examinateur*
Pier-Giorgio ZANONE - Université Paul Sabatier, Toulouse	*Examinateur*
Didier BAZALGETTE – MRIS, DGA	*Examinateur*
Simon Thorpe - CNRS, Université Paul Sabatier, Toulouse	*Directeur de thèse*

Ecole doctorale : *CLESCO*
Unité de recherche : *Centre de Recherche Cerveau et Cognition, UMR 5549*
Directeur(s) de Thèse : *Simon J. Thorpe*

Remerciements

À Simon Thorpe, qui m'a encadré depuis mes débuts dans la recherche en Master jusqu'à la fin de cette thèse. Cela a toujours été un plaisir de travailler avec toi, tes idées foisonnantes et ton enthousiasme perpétuel ont été pour moi un moteur important tout au long de ces quatre années. Même si tu es très occupé, tu as toujours su exploiter au maximum tes moments au laboratoire pour mettre le doigt sur ce qui était important dans nos recherches. J'espère avoir appris de toi cette capacité à voir l'essentiel, et surtout à toujours essayer de faire de la recherche originale.

À Françoise Vitu-Thibault et Guillaume Rousselet, qui ont accepté d'être rapporteurs de ce travail et l'ont enrichi par leur remarques pertinentes. Je remercie aussi Olivier Pascalis, Bruno Rossion, Didier Bazalgette et Pier-Giorgio Zanone d'avoir accepté de participer au jury de ma thèse.

À Holle Kirchner. Tu m'as guidé dans mes premiers pas expérimentaux, tu m'as appris les fondamentaux de la psychophysique et la rigueur qu'elle implique. Je t'en serai éternellement reconnaissant. J'espère avoir été un bon élève.

À tous les membres du CerCo. D'abord Michèle qui réussit tous les jours à instaurer une ambiance studieuse et agréable au sein du laboratoire. Ensuite Denis, Rufin, Emmanuel, Lionel, Jean-Michel et Leila qui ont toujours su apporter des réponses précises à mes questions ou me guider vers les bonnes références. Je tiens aussi à remercier l'ensemble des personnels techniques et administratifs du CerCo qui se sont toujours rendus disponibles pour me rendre la vie plus facile. Enfin, un immense merci à Nadège et Olivier d'avoir pris le temps de relire et corriger ce mémoire dans son intégralité.

À tous les étudiants de Psychomotricité et de Psychologie que j'ai eu la chance d'avoir dans mes classes de TD. Leur gentillesse (et leur indulgence au début) m'a donné goût à l'enseignement. Je les remercie aussi pour leur importante participation à mes expériences en tant que cobaye. J'en profite au passage pour remercier toutes les personnes qui ont accepté de participer à mes expériences.

À tous ceux qui sont passés dans le bureau des étudiants du bâtiment A3. Celui-ci a été un lieu de travail et de discussion génial durant mes quatre années ici (il était temps que je m'en aille, je commençais à être le dinosaure du bureau). Je tiens à remercier tous ceux qui y sont passés, au début Ludovic, Julien et Olivier, puis ensuite Vince, Leila, Evelyne et Nadège. Mes questions, quelles qu'elles soient, ont toujours trouvé des réponses et votre présence a rendu la vie quotidienne au laboratoire des plus agréables.

À tous ceux qui tentaient par tous les moyens de m'empêcher de travailler en allant au squash, au wakeboard ou en faisant un volley en bas du CerCo : Rufin, Ludo, Oliv, Julien, Gabriel, Max, Adrien et Romain.

À tous les compagnons de pause cigarette, de sorties ou de discussions au ru : Marianne, Oliv, Max, Yanica, Romain, Gladys, Gabriel, Rama, Marlène et aussi Aymeric.

Je tiens aussi à remercier mes parents et ma petite sœur pour leur soutien constant et leur confiance, ainsi que mon frère pour ses conseils avisés.

Et bien sûr, je remercie infiniment Tévy. Ta patience, ton soutien et ton amour ont été la base de ma réussite dans ce long projet. Je ne te serai jamais assez reconnaissant pour tout ce que tu as fait pour moi.

À Tévy,

Publications

Article publié

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces : Face detection in just 100 ms. *Journal of Vision*, 10(4) :16, 1-17, http ://journalofvision.org/10/4/16/, doi :10.1167/10.4.16.

Articles en préparation

Crouzet, S.M. & Thorpe, S.J. Swap the face ! Use of amplitude spectrum information to drive ultra-fast saccades.

Crouzet, S.M., Kirchner, H., Bayerl, P., Neuman, H. & Thorpe, S.J. Processing times for optic flow patterns measured by the saccadic choice task.

Crouzet, S.M., Macé, M., Bacon-Macé, N., Fabre-Thorpe, M. & Thorpe, S.J. Masking in a high-level gender discrimination task is essentially entirely pre-cortical.

Wu, C.T., Crouzet, S.M., Thorpe, S.J. & Fabre-Thorpe, M. At 120 ms you know where the animal is but you don't yet know it's a dog.

Joubert*, O.R., Crouzet*, S.M., Thorpe, S.J. & Fabre-Thorpe M. Timing the earliest object-context interactions.
* Auteurs à égales contributions

Chapitre en français

Fabre-Thorpe, M., Crouzet, S., Rousselet, G. A., Kirchner, H., & Thorpe, S. J. (2008). Catégorisation visuelle rapide : les visages sont-ils des objets spécifiques ? In *Traitement et reconnaissance des visages : du percept à la personne*. E. J. Barbeau, S. Joubert and O. Felician. Marseille, Solal : 239-260.

Résumés de conférences publiés

Simon J. Thorpe, Adrien Brilhault & Sébastien M. Crouzet (soumis) Colour based target selection for ultra-rapid saccades : The fastest controllable selection mechanism ? ECVP 2010, Lausanne, Switzerland.

Marie Mathey, Sébastien M. Crouzet & Simon J. Thorpe (soumis) The accuracy of ultra-rapid saccades to faces. ECVP 2010, Lausanne, Switzerland.

Crouzet, S. M. & Thorpe, S. J. (2010) Power spectrum cues underlying ultra-fast saccades towards faces. VSS, Naples, Florida.

Mathey, M. A., Crouzet, S. M. & Thorpe, S. J. (2010) Ultra-rapid saccades to faces : the effect of target size. VSS, Naples, Florida.

Crouzet S, Mathey M & Thorpe S J (2009). Ultra-fast saccades to faces : A temporal precedence effect ? Perception 38 ECVP Abstract Supplement, page 157.

Crouzet, S. M., Joubert, O. R., Thorpe, S. J., & Fabre-Thorpe, M. (2009). The bear before the forest, but the city before the cars : Revealing early object/background processing [Abstract]. Journal of Vision, 9(8) :954, 954a, http ://journalofvision.org/9/8/954/, doi :10.1167/9.8.954.

Fabre-Thorpe, M., Crouzet, S. M., Wu, C.-T., & Thorpe, S. J. (2009). At 130 ms you "know" where the animal is but you don't yet "know" it's a dog [Abstract]. Journal of Vision, 9(8) :786, 786a, http ://journalofvision.org/9/8/786/, doi :10.1167/9.8.786.

Thorpe, S. J., Crouzet, S. M., Macé, M. J., Bacon-Macé, N., & Fabre-Thorpe, M. (2009). Masking in a high-level gender discrimination task is essentially entirely pre-cortical [Abstract]. Journal of Vision, 9(8) :546, 546a, http ://journalofvision.org/9/8/546/, doi :10.1167/9.8.546.

S Crouzet, H Kirchner, S J Thorpe (2008). Saccading towards faces in 100 ms. What's the secret ? Perception 37 ECVP Abstract Supplement, page 119.

S J Thorpe, H Kirchner, S Crouzet, P Bayerl, H Neumann (2008). Processing times for optic flow patterns measured by the saccadic choice task. Perception 37 ECVP Abstract Supplement, page 40.

Crouzet, S., Thorpe, S. J., & Kirchner, H. (2007). Category-dependent variations in visual processing time. Journal of Vision, 7(9) :922,922a, http ://journalofvision.org/7/9/922/, doi :10.1167/7.9.922.

Thorpe, S., Crouzet, S., & Kirchner, H. (2007). Saliency maps and ultra-rapid choice saccade tasks. Journal of Vision, 7(9) :30, 30a, http ://journalofvision.org/7/9/30/, doi :10.1167/7.9.30.

Simon J. Thorpe, Sébastien Crouzet, Holle Kirchner and Michèle Fabre-Thorpe (2006). Ultra-rapid face detection in natural images : implications for computation in the visual system. First French Conference on Computational Neurosciences, pp. 124-127. Abbaye des Prémontrés, Pont à Mousson, France.

Simon J. Thorpe, Sébastien Crouzet and Holle Kirchner (2006). Comparing processing speed for complex natural scenes and simple visual forms. Perception, vol. 35, p 128.

Préface

Les ordinateurs peuvent aujourd'hui réaliser de nombreuses tâches plus rapidement et plus précisément que les humains : calculer $\sqrt{3}$, retrouver quel jour de la semaine était le 15 juin 1975, prévoir vingt coups à l'avance aux échecs. Pourtant, ils sont encore loin de pouvoir rivaliser avec nos performances dans certaines tâches qui peuvent nous sembler triviales. Il est ainsi tout a fait possible de programmer un robot pour qu'il se déplace correctement sur une surface plane. Par contre, posez le sur un trottoir de centre ville et il ne tiendra pas une minute. Le monde réel est *bruité* et *complexe*, et pour l'instant les robots ont beaucoup de difficulté à gérer ce type de situation.

Un autre exemple dans lequel nous écrasons largement les ordinateurs, et ceci dès notre plus jeune âge, est la reconnaissance d'objets. Malgré la complexité de cette tâche dans un environnement naturel, nous la réalisons sans nous en rendre compte des centaines de fois chaque jour, la plupart du temps inconsciemment et en un temps record (personne n'a jamais attendu que son système visuel finisse l'analyse, cela se passe de façon *immédiate*). En plus de sa vitesse et de son automaticité, le système visuel humain réussi à atteindre une très bonne sélectivité (différencier facilement vos deux fils, même s'ils sont jumeaux) tout en gardant une grande robustesse (reconnaître un ami d'école 30 ans après de façon quasi-immédiate malgré ses cheveux absents et ses 30 kg de plus). Aucun système informatique n'a pour le moment réussi à concilier ces deux capacités comme le fait le système visuel humain (Riesenhuber et Poggio, 1999).

Cette thèse s'inscrit donc dans le projet d'étude des mécanismes permettant de tel performances. Plus précisément, j'ai rejoins le groupe de Simon Thorpe au Centre de Recherche Cerveau et Cognition pour étudier spécifiquement la dynamique de processus mis en jeu lors de la perception de scènes naturelles. Elles seraient en effet un médium plus "naturel" pour le système visuel et permettraient donc de le tester dans des conditions plus réalistes. L'objectif de cette équipe, depuis maintenant plusieurs décennies, est de comprendre les mécanismes cérébraux nous permettant de reconnaître en un temps record la présence d'objet (animaux, visages, véhicules) dans des scènes complexes. Ce temps minimum était supposé être d'environ 150 ms, selon les études précédentes en éléctroencéphalographie. Cependant, comme nous le verrons au cours de ce travail, il pourrait être encore bien plus court, car nous pourrions produire des réponses comportementales bien avant ces 150 ms !

Pour commencer, je dresserai un état de l'art détaillé du domaine. Nous verrons comment le cerveau commence par traiter les informations visuelles, les hypothèses majeures sur l'utilisation de ces informations pour reconnaître les objets, puis les contraintes temporelles imposé à ces modèles par les travaux sur la perception rapide. Ensuite, nous verrons comment les mouvements oculaires peuvent nous permettre d'aller encore plus loin en ayant pris soin de dresser un résumé précis de leur fonctionnement. Les deuxièmes et troisièmes chapitres présenteront un certain nombre d'études expérimentales réalisées durant ma thèse. Je présenterai tout d'abord une étude qui a permis de montrer des variations de temps de traitement selon la catégorie d'objet, et surtout un avantage clair pour une catégorie spécifique : les visages. Une deuxième étude permettra de préciser la nature de l'information à la base de cet avantage. Dans une troisième étude, nous avons étudié le rôle spécifique des informations contextuelles dans la reconnaissance rapide d'objets. Une dernière étude expérimentale permettra d'aller plus loin dans la compréhension de la nature des représentations accessibles si précocement, en montrant que celles-ci pourraient être plutôt rudimentaires. Je me servirai ensuite de l'ensemble de ces résultats pour proposer, dans le quatrième chapitre, un modèle simple de décision saccadique permettant d'expliquer, par exemple, l'avantage des visages au niveau comportemental. Enfin, je terminerai par une discussion de l'ensemble de ces résultats expérimentaux dans le cinquième chapitre.

Table des matières

Chapitre 1

La perception visuelle

Sommaire

1.1 L'objet : les scènes naturelles

Le système visuel est un système perceptif apparu progressivement au cours de l'évolution pour nous permettre d'utiliser les variations d'intensité lumineuse de l'environnement. Pour comprendre son fonctionnement, il est donc primordial d'explorer la structure de l'information qu'il traite. Gibson avait par exemple très tôt envisagé le problème dans ce sens, en s'intéressant à la structure du flux visuel et à ses variations,

et en le mettant au centre de sa théorie de la perception visuelle (Gibson, 1979). L'hypothèse est que le système visuel s'est construit de manière à exploiter les régularités de son environnement, ce que l'on peut appeler les contraintes écologiques. Comprendre celles-ci constitue donc une approche complémentaire (voir préalable) à l'étude du système perceptif en lui-même (Geisler, 2008). Concernant l'étude du système visuel, et plus précisément de la reconnaissance d'objets, les stimuli les plus utilisés sont souvent des photographies. Même si elles ne sont pas une représentation exacte du monde réel (seulement en deux dimensions, figées et limitées dans l'espace), elles permettent tout de même d'étudier le système visuel dans des conditions assez proches de la réalité.

1.1.1 Validité écologique vs. contrôle des paramètres

L'utilisation des photographies est aujourd'hui loin d'être la norme en science de la vision. En effet, l'immense majorité des chercheurs utilisent des stimuli artificiels simples, partant du principe que, même si ces stimuli ne sont pas naturels, ils permettent tout de même de comprendre certains mécanismes précis tout en contrôlant parfaitement les paramètres expérimentaux (Rust et Movshon, 2005). Cette approche a incontestablement permis, depuis une cinquantaine d'années, des avancées considérables dans les neurosciences de la vision. Cependant, certains résultats expérimentaux viennent aujourd'hui complexifier la tâche, en montrant que ce qui est vrai avec des stimuli artificiels ne l'est pas forcément avec des scènes naturelles. Par exemple, une étude comportementale utilisant un protocole de double tâche a montré que l'attention était nécessaire pour dire si la lettre présentée était un T ou un L, mais pas pour dire si l'image contenait ou non un animal (Li *et al.*, 2002). Alors que l'on aurait pu imaginer que reconnaitre des lettres était beaucoup plus simple que détecter la présence d'un animal dans une scène naturelle (ce qui est vrai pour un ordinateur), c'est l'inverse qui se produit ici. Les processus différent donc selon que l'on utilise des stimuli artificiels ou des scènes naturelles. Il faut donc être particulièrement vigilant avant de généraliser les résultats expérimentaux obtenus en laboratoire.

De nombreux autres exemples nous viennent de l'électrophysiologie. Depuis une vingtaine d'années, de nombreuses études ont étudié les différences d'activité neuronale dans différentes aires visuelles, selon que l'on présentait des stimuli artificiels ou naturels. La région la plus étudiée a été le cortex visuel primaire (V1), porte d'entrée de l'information visuelle vers le cortex. La réponse des neurones de cette région est différente pour des stimuli artificiels et pour des stimuli naturels (Kayser *et al.*, 2003). En réponse à une image naturelle, les neurones ont tendance à produire plus de décharges par paquets (anglais : burst) et à avoir une activité globale moins importante (Gallant *et al.*, 1998). Le patron de réponse globale des neurones de V1 est donc plus épars (anglais : sparse) et intense (Graham et Field, 2007). Pris dans leur ensemble,

ces résultats montrent que lorsque l'on présente des scènes naturelles, l'information par décharge augmente pour chaque neurone, le codage est donc plus efficace que pour des stimuli artificiels (Vinje et Gallant, 2000). Le codage de l'information visuelle serait donc optimisé pour le traitement des scènes naturelles (Simoncelli et Olshausen, 2001).

Cependant, le principal reproche qui est fait à l'utilisation des scènes naturelles en science de la vision est le manque de contrôle sur les paramètres que l'on veut faire varier (Rust et Movshon, 2005). Le paradigme classique en sciences expérimentales est le "toutes conditions égales par ailleurs" : faire varier le minimum de paramètres pour être certain que l'effet observé est bien causé par la variable que l'on manipule. Le risque avec les scènes naturelles est alors que l'effet observé soit causé par un biais dans les images n'ayant rien à voir avec la question expérimentale. Deux études récentes ont illustré ce risque en montrant que des modèles simples pouvaient avoir des performances tout à fait respectables dans une tâche de catégorisation en exploitant les variations du fond ou les biais des groupes d'images (Pinto *et al.*, 2008; Wichmann *et al.*, 2010). Pour compenser au maximum ce risque, il est donc important de bien comprendre la structure et les paramètres des images, ce que vont permettre dans une certaine mesure les statistiques. Ainsi, même si le paramétrage précis des scènes naturelles et encore loin d'être abouti, elles semblent être un bon compromis entre validité écologique (le monde réel) et contrôle des paramètres (stimuli artificiels) (Felsen et Dan, 2005; Kayser, 2004).

1.1.2 Statistiques d'images

Contrairement à ce que l'on pourrait penser intuitivement, les images naturelles ne sont qu'une infime fraction de l'ensemble des images que l'on peut construire en deux dimensions. Par exemple, si l'on attribue une valeur aléatoire à chaque pixel d'une image, celle-ci ne ressemblera certainement pas à ce que l'on peut rencontrer dans notre environnement (cette image ressemblera plus probablement à du bruit blanc).

Distribution des pixels

Une première façon d'étudier les statistiques des images naturelles est de s'intéresser à la distribution de la luminance des pixels au sein d'une image. Ces distributions peuvent être très variables entre les images. On les résume souvent par leur valeur moyenne et leur écart-type. Il est important de noter que ces valeurs ne rendent pas compte de toutes formes possible de distributions. Ainsi, deux images peuvent avoir la même luminance moyenne et le même écart-type, tout en ayant des distributions assez différentes, par exemple dans le cas où l'une des deux images a une distribution bi-modale et pas l'autre. Ces paramètres étant faciles à contrôler, la plupart des études utilisant des scènes naturelles prennent soin de les égaliser entre les images.

Les fréquences spatiales

Comme nous l'avons déjà évoqué dans la section 1.1.1, les filtres spatiaux semblent être un bon modèle du système visuel, au moins pour les premières étapes. En effet, de nombreux chercheurs ont depuis longtemps assimilé les opérations réalisées lors de ces premières étapes à une transformée de Fourier (Westheimer, 2001). Les outils mathématiques élaborés dans ce cadre vont donc fournir des informations importantes pour l'analyse des statistiques des images naturelles.

Selon la transformée de Fourier, tout signal complexe peut être décomposé en une "somme infinie" de fonctions sinusoïdales à différentes fréquences. Quand on l'applique à une image en deux dimensions, la transformée de Fourier permet de la décomposer en une somme de fréquences spatiales ayant des fréquences et des amplitudes différentes. Ceci peut être réalisé sur l'ensemble de l'image, ou alors sélectivement pour chaque orientation (verticale, horizontale, ou obliques).

Chaque fréquence composant ce spectre peut alors être décrite selon deux caractéristiques : son amplitude et sa phase. L'amplitude correspond à l'intensité de sa contribution dans l'image donnée. Si une fréquence donnée n'est pas présente dans le spectre d'une image, son amplitude sera de 0. Le spectre de phase régit quant à lui l'organisation de l'ensemble des fréquences du spectre dans l'image, et va donc particulièrement compter pour l'apparition des bords dans une image (Oppenheim et Lim, 1981; Piotrowski et Campbell, 1982). Par exemple, pour un endroit donné dans l'image, si plusieurs fréquences du spectre sont en phase, leurs amplitudes vont s'additionner et un bord va apparaître. La phase est donc la principale responsable des contours de l'image, et permet d'en reconnaître les détails, comme le montre la figure 1.1.

Même si le spectre d'amplitude semble jouer un rôle au mieux marginal pour le "sens" de l'image, il reste très intéressant pour décrire statistiquement les images. Ainsi, on retrouve une distribution similaire des fréquences spatiales au sein de l'ensemble des scènes naturelles, avec des basses fréquences sur-représentées. Plus on monte dans les fréquences, moins leur amplitude devient importante. Ceci résulte en une distribution en $\frac{1}{f^n}$ qui est la caractéristique principale du spectre d'amplitude des scènes naturelles. Cependant, le spectre d'amplitude pourrait contenir de l'information sur la catégorie de l'image, même si ceci n'est pas manifeste au niveau perceptif. Ainsi Aude Oliva et Antonio Torralba ont montré dans plusieurs études que le spectre d'amplitude pouvait être utilisé par un ordinateur pour classifier les scènes selon leur contenu (Oliva et Torralba, 2001, 2006; Torralba et Oliva, 2003). Il semble aussi que des images artificielles ne contenant que des informations d'amplitude peuvent biaiser la réponse des sujets lorsqu'il s'agit de catégoriser des images (Kaping *et al.*, 2007).

Ce qui est fascinant dans cette approche est qu'elle a trouvé de nombreux échos dans le traitement que font les neurones eux-même, qui peuvent être assimilés indivi-

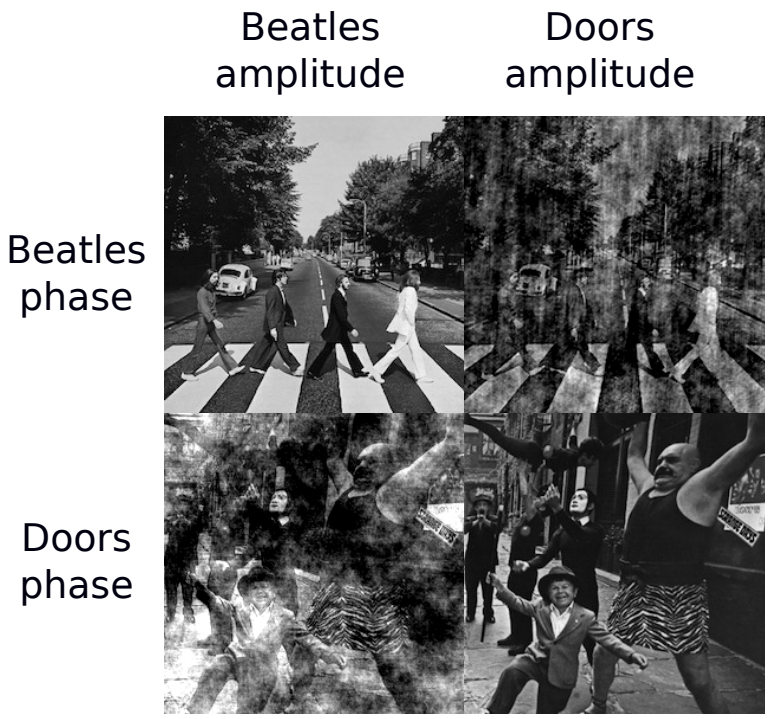

FIGURE 1.1: Démonstration de comment la phase et l'amplitude influence l'apparence des images. Les images originales ont été transformées en noir et blanc avant les traitements dans le domaine fréquentiel. L'image originale en haut à gauche est celle de l'album Abbey Road des Beatles, celle en bas à droite est la pochette de l'album Strange Days des Doors.

duellement à des filtres spatiaux. Et même si le rôle du spectre de phase semble être crucial pour la reconnaissance d'objets, le spectre d'amplitude pourraient aussi avoir un rôle non négligeable. Une des expériences de cette thèse teste ainsi le rôle effectif des informations d'amplitude pour la détection saccadique rapide d'objet dans le champ visuel.

1.2 Les premières étapes de la perception

La perception de notre environnement visuel est basée sur la lumière. Lorsque celle-ci, reflétée par les objets qui nous entourent, atteint l'œil, la cornée et le cristallin vont jouer le même rôle que l'objectif d'un appareil photo. La lumière les traversant va alors former une image inversée sur la rétine qui tapisse le fond de l'œil. Les photons composant la lumière vont alors être captés par des cellules spécifiques de la rétine : les photorécepteurs. On en trouve deux types : les cônes et les bâtonnets. Les bâtonnets, plus nombreux, fonctionnent dans des conditions de luminosité réduite, alors que les cônes fonctionnent lorsque la luminosité est normale. La concentration des cônes est maximale au centre de la rétine, appelé fovéa, où l'acuité visuelle va être à son maximum. Plus l'on s'éloigne du centre de la rétine, plus celle-ci diminue. La répartition des photorécepteurs n'est donc pas uniforme, et assure une définition maximale uni-

quement au centre du champ visuel. Cette limite est compensée par les mouvements oculaires, qui permettent à chaque instant d'amener sur la fovéa la portion du champ visuel qui nous intéresse.

La rétine dans son ensemble est un mini système nerveux (Gollisch et Meister, 2010), constitué d'une grande variété de cellules. Pour simplifier, on considère généralement qu'elle est organisée en trois couches distinctes constituées par différents types de cellules : les photorécepteurs qui constituent la couche d'entrée, les cellules bipolaires, et enfin les cellules ganglionnaires. Au sein même de ces couches, d'autres types de cellules vont assurer le rôle de connexion horizontales (par opposition à l'organisation verticale entre les couches). Cette organisation en couche constitue une signature du système visuel dans sa globalité, et permet déjà un pré-traitement du signal visuel. Ainsi, l'image en deux dimensions du champ visuel qui s'imprime sur les photorécepteurs constitue une représentation matricielle optimale du champ visuel, avec environ 130 millions de photorécepteurs (par analogie avec les images numériques, ceux-ci correspondraient à des pixels). À sa sortie de la rétine, le nombre de "pixels" a chuté, l'image n'étant plus codée que par 1 million de cellules ganglionnaires. Cependant, la géométrie des connexions entre les neurones de chaque couche aboutit en une organisation en champ récepteur. Cette organisation, pour continuer l'analogie, fait que cette chute du nombre de "pixels" correspond plus à une compression qu'à une perte d'information.

Le champ récepteur d'un neurone est la région du champ visuel qui, lorsqu'elle est stimulée de façon appropriée, module sa réponse.

Une cellule ganglionnaire donnée aura de ce fait un champ récepteur directement défini par l'ensemble des photorécepteurs qui lui fournissent de l'information. La cellule ganglionnaire ne va pas simplement sommer les activités de ces cellules afférentes mais déjà faire une opération de détection de bords. En effet, si on prend l'exemple d'une cellule ganglionnaire centre ON / pourtour OFF, les cellules afférentes lui fournissant de l'information sur le centre du champ récepteur l'exciteront. Au contraire, les cellules lui fournissant de l'information sur le pourtour de son champ récepteur l'inhiberont. Cette configuration fait que la cellule ganglionnaire cible déchargera peu sur une surface uniforme (qui stimulera de la même façon son centre et son pourtour) mais s'activera beaucoup lorsqu'elle tombera sur un bord de l'image.

L'ensemble des axones des cellules ganglionnaires va ensuite sortir du globe oculaire via le nerf optique. Celui-ci va se projeter très majoritairement vers les corps genouillés latéraux (CGL) du thalamus (une partie des autres cibles du nerf optique sera abordée dans la section 2.2.4). Longtemps, le CGL n'a été considéré que comme un simple relai car les champs récepteurs de ses neurones ne différaient pas significativement de ceux de la rétine. Cependant, ce qui le rend extrêmement intéressant est le fait qu'une grande partie de ses fibres afférentes ne proviennent pas de la rétine mais du cortex. Il pourrait donc jouer un rôle fondamental de modulation du signal provenant de la

rétine (McAlonan *et al.*, 2008; Rees, 2009; Crick, 1984). Son fonctionnement, au moins chez l'animal anesthésié (ce qui supprime certainement ces modulations complexes) commence en tout cas à être très bien modélisé (Mante *et al.*, 2008; Reinagel *et al.*, 1999; Lesica et Stanley, 2004).

Après cette nouvelle étape de traitement, l'information est acheminée vers le cortex, où elle entrera via la couche IVc β du cortex visuel primaire (V1). La découverte des caractéristiques des champ récepteur de cette région par Hubel et Wiesel constitue certainement l'acte fondateur du modèle actuel du système visuel (Hubel et Wiesel, 1959). En enregistrant les neurones de cette aire chez le chat anesthésié, ils ont ainsi montré qu'après les champs récepteurs de type ON/OFF circulaire du LGN, les neurones de V1 étaient particulièrement sélectifs à des formes de barre orientée. Encore une fois, une géométrie précise permet de faire émerger ce type de champ récepteur. En effet, il suffit de connecter un neurone de V1 à une série de neurone du LGN (ceux-ci ayant des champs récepteurs circulaires), selon une organisation bien précise (ici en ligne) pour faire émerger un champ récepteur sélectif aux barres orientées. C'est donc encore ici la géométrie des connexions qui structure l'aspect fonctionnel de ces neurones. Plus précisément, le type de champ récepteur que je viens de présenter correspond aux cellules simples de V1, mais on y trouve aussi d'autres types (complexes, hyper complexes) qui permettent des traitements plus élaborés. Plus que la découverte en soi, c'est surtout sa compatibilité avec les théories du système visuel comme un système de filtre qui en fait une des études fondatrices des neurosciences actuelles de la vision. Comme nous l'avons vu dans le chapitre précédent (voir section 1.1.1), ce type de champ récepteur s'accorde parfaitement avec une conception du système visuel comme analyseur de Fourier ou à ondelettes.

L'organisation en couches est donc un aspect fondamental des premières étapes du système visuel. Le nombre de neurones "représentants" le champ visuel diminue à chaque étape, et c'est la géométrie des connexions entre les couches qui va définir le traitement opéré. Ainsi, chaque neurone fait un travail similaire : il décharge en fonction de l'activité enregistrée au niveau de ses synapses. Mais la géométrie bien spécifique des connexions entre les neurones permet de créer des détecteurs de contraste dans la rétine, puis de barres orientées dans le cortex visuel primaire (V1). Si l'on ne considère pour l'instant que la traitement ascendant de la rétine vers le cortex, la fonction d'un neurone tiré au hasard dans la voie visuelle est donc entièrement définie par l'arrangement de l'ensemble des neurones qui participent à ses afférences (ce qu'on appelle son champ récepteur).

1.3 Reconnaître les objets

Comme nous l'avons vu dans le chapitre précédent (voir chapitre 1.2), V1 constitue l'entrée principale de l'information visuelle dans le cortex visuel. Ensuite, le tableau se complexifie énormément, mais certains grands principes peuvent être extraits de cette complexité. Le premier est la *spécialisation fonctionnelle* qu'on peut traduire en langage courant par division du travail.

Spécialisation fonctionnelle : Des voies neuronales spécialisées traitent les différents aspects de la scène visuelle : par exemple la couleur, la forme, ou le mouvement.

Après V2, les informations visuelles vont donc se ségréguer en deux voies distinctes : les voies ventrale et dorsale, chacune ayant des rôles différents (Mishkin *et al.*, 1983). La voie ventrale est ainsi généralement considérée comme le lieu de la reconnaissance d'objets, elle correspond aux connexions de V1 au cortex inférotemporal (IT) en passant par V4. La voie dorsale quant à elle extrairait les informations visuelles nécessaires pour orienter les mouvements oculaires, guider la prise des objets, ou toute tâche nécessitant une localisation. Elle correspond aux projections des neurones de V2 vers V3 puis MT (aussi appelé V5) avant d'atteindre des zones comme 7A et IP (intrapariétal). Nous nous intéresserons dans cette section aux modèles qui ont été proposés pour expliquer les mécanismes permettant la reconnaissance d'objets. La plupart des références anatomiques se rapporteront donc à des aires de la voie ventrale. Cependant, comme nous le verrons en fin de chapitre, certaines aires du cortex pariétal, et donc généralement associées à la voie dorsale, pourraient aussi participer à la reconnaissance d'objets.

Tout au long de la voie ventrale, l'organisation rétinotopique, évidente dans V1 et V2, va s'estomper progressivement dans V4 et IT. Les champs récepteurs quand à eux vont très largement s'agrandir : par rapport à la rétine, ils sont 30 fois plus large dans V4, et jusqu'à 100 fois plus large dans IT. Ceci illustre directement le deuxième grand principe déjà abordé dans la section 1.2 : le *traitement hiérarchique*.

Traitement hiérarchique : La perception visuelle se fait par des traitements graduels dans lesquels l'information est d'abord représentée sous forme simple et localisée, puis transformée étape après étape en une information plus abstraite, holistique, voir même multimodale (ce concept est illustré dans la figure 1.2).

1.3.1 Représentation intermédiaire : aire V4

Comme l'illustre la figure 1.2, l'hypothèse hiérarchique voudrait qu'après des détecteurs d'orientation dans V1, on trouve des neurones sélectifs à des combinaisons de traits dans les aires suivantes, jusqu'aux neurones sélectifs à des objets dans les aires de haut-niveau comme IT. L'aire V4, qui se trouve entre les deux, doit alors jouer un

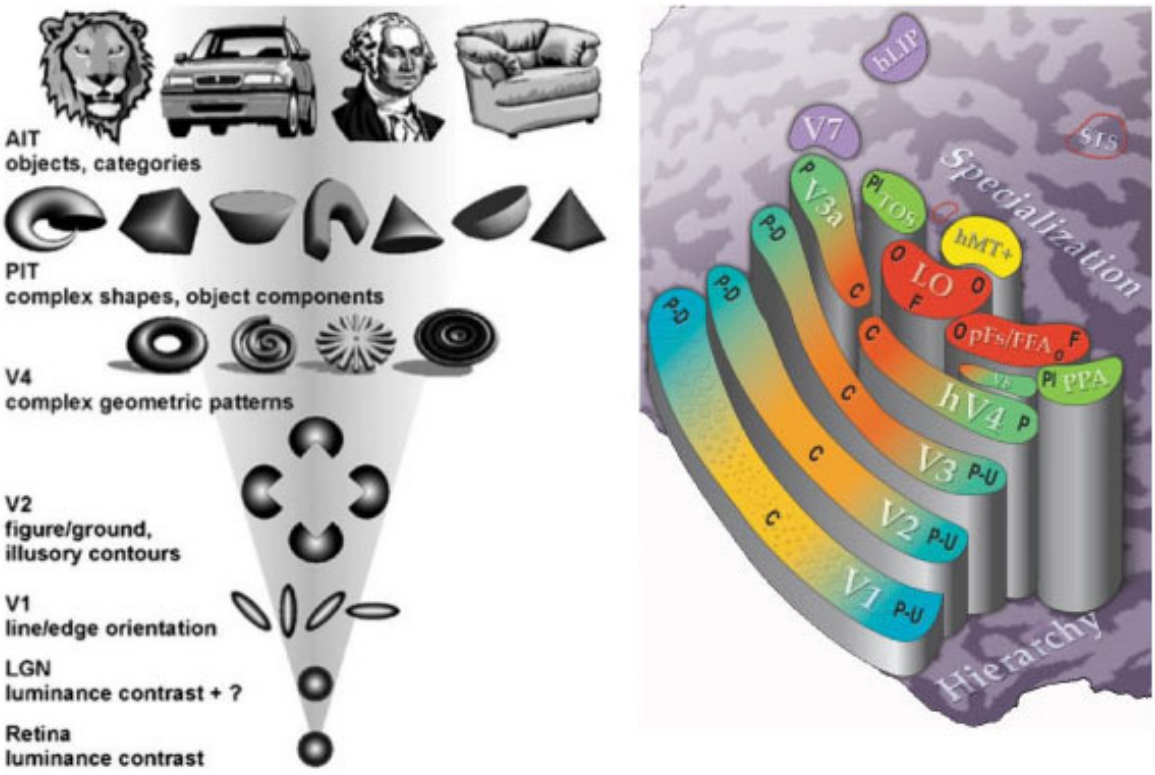

FIGURE 1.2: **Figure de gauche :** Schéma de l'organisation hiérarchique des champ récepteur dans la voie ventrale du système visuel humain. Ceux-ci, circulaire dans la rétine et le LGN, deviennent orientés dans V1. Dans V2, certains neurones s'activeraient pour la colinéarité par exemple, ce qui expliquerait l'apparition des contours illusoires. Puis au sommet, on trouve des neurones dans AIT qui répondent sélectivement à des objets précis (visages, véhicules, animaux). Il n'y a cependant pas de consensus sur ce que seraient les représentations intermédiaires contenues dans les neurones de V4 ou de PIT. Extrait de VanRullen (2003). **Figure de droite :** Atlas schématique du cortex visuel. Chaque aire est représentée ici dans l'hémisphère droit. Les aires sont représentées en escalier, plus elles sont hautes dans l'escalier, plus elles sont hautes dans la hiérarchie visuelle. Extrait de Grill-Spector et Malach (2004).

rôle crucial en transformant des traits physiques simples en représentations abstraites. Les neurones de V4 devraient donc être sélectifs à des représentations intermédiaires, plus évoluées que dans V1 mais pas encore au même niveau d'abstraction que dans IT. Ce type de représentations intermédiaires aurait un rôle clé pour la reconnaissance d'objets (Ullman *et al.*, 2002; Ullman, 2007). Il semble en effet qu'une lésion dans V4 entraîne une dégradation des performances dans des tâches de discrimination de formes (Schiller et Lee, 1991). D'autre part, les enregistrements cellulaires dans V4 soulignent une sélectivité importante à la forme, que ce soit pour des barres ou gratings orientés (Desimone et Schein, 1987), ou pour des stimuli plus complexes comme des cercles concentriques ou des spirales (Gallant *et al.*, 1996). Par ailleurs, V4 aurait, comme V1, une organisation modulaire avec une organisation en colonnes des neurones codant pour des propriétés similaires (Ghose et Ts'O, 1997).

La représentation la plus aboutie, pour l'instant, des caractéristiques des champs récepteurs de V4 est certainement celle issue des travaux de l'équipe de Jack Gallant. Ils ont ainsi proposé que la sélectivité des neurones de V4 correspondrait à un champ récepteur spectral (anglais : Spectral Receptive Field, SRF). Cette description, dans le domaine fréquentiel et non-plus celui de l'image, permet d'expliquer les résultats très disparates concernant les stimuli auxquels un même neurone de V4 peut être sélectif. En utilisant une méthode de corrélation inverse et des stimuli "naturels", ils ont démontré que chaque neurone de cette aire serait ainsi sélectif à différentes orientations et à différentes fréquences spatiales (David *et al.*, 2006). De plus, il semblerait que cette sélectivité des neurones de V4 puisse être modulée par la tâche à réaliser (David *et al.*, 2008). Il est important de remarquer que l'immense majorité des découvertee sur l'aire V4 ont été faites chez le singe. Il semble cependant qu'on trouve une aire similaire chez l'homme (Gallant *et al.*, 2000; Hansen *et al.*, 2007) à laquelle il est le plus souvent fait référence sous le nom de V4v.

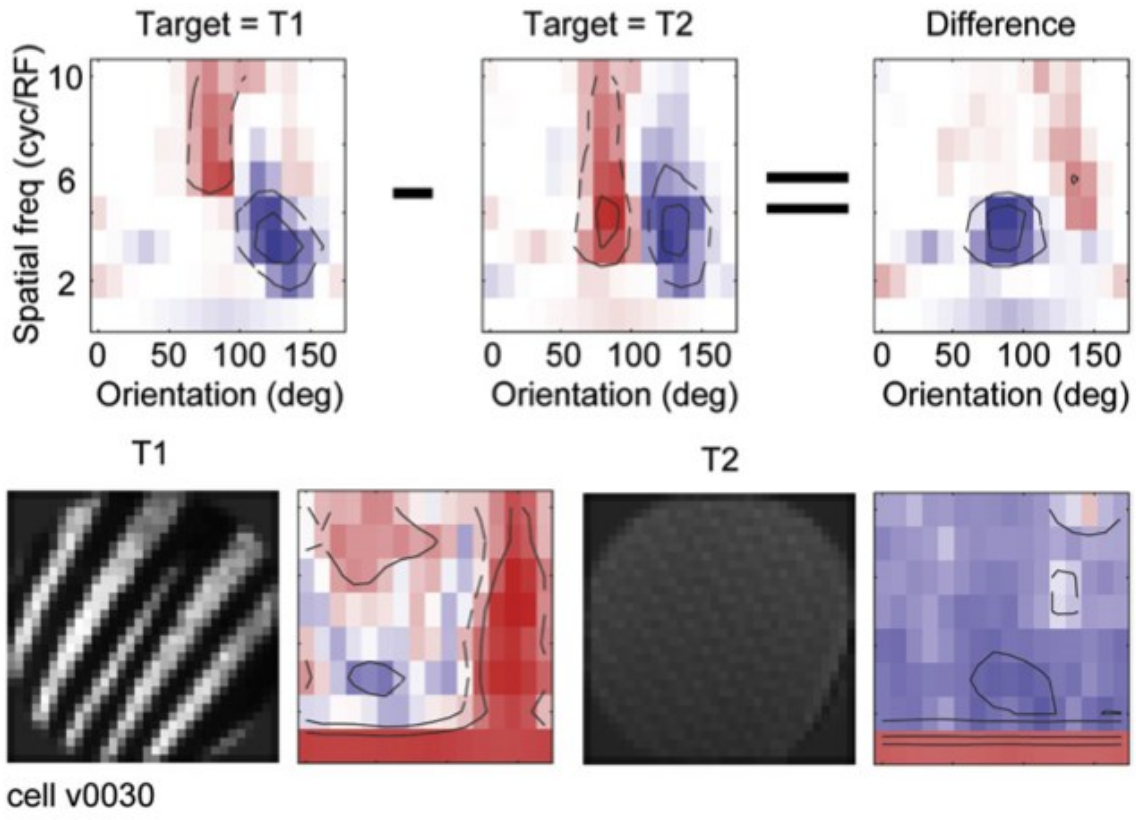

FIGURE 1.3: Modulation de la sélectivité des neurones de V4 par la tâche (David *et al.*, 2008). Chaque graphique représente la sélectivité d'un même neurone de V4 sous forme de champ récepteur spectral (SRF). Celui-ci va changer les caractéristiques auxquelles il est sélectif selon que le singe cherche la cible T1 ou la cible T2.

1.3.2 Cortex occipital latéral et inféro-temporal

Dans V4, le champ visuel est encore représenté selon une organisation rétinotopique. Celle-ci se perd ensuite au profit d'une plus grand invariance à la taille et à la position des objets par exemple. Avant de passer en revue les données électrophysiologiques, obtenues très majoritairement chez le singe macaque, nous allons d'abord nous intéresser aux recherches qui ont été menées chez l'homme (Grill-Spector et Malach, 2004).

Pour localiser les régions contenant des neurones directement impliqués dans le traitement des objets, le groupe de Malach a testé la réponse des aires du cortex visuel lorsque l'on présentait des images ayant une structure spatiale plus ou moins "mélangée". L'expérience consistait à diviser les images en grille puis à mélanger les carrés produits, plus les carrés étaient petits, plus le mélange détruisait la structure spatiale de l'image et lui faisait perdre son sens. Ils observaient ensuite la réponse globale en IRMf de différentes aires en fonction de l'intensité du mélange (voir figure 1.4). Les régions répondant plus aux images intactes qu'aux images "mélangées" pouvaient ainsi être associées au traitement spécifique des objets, et ont été divisées en trois groupes selon leur localisation (Grill-Spector *et al.*, 1998) : le Complexe Occipital Latéral (LOC), la région occipito-temporale ventrale (VOT), ainsi que certaines régions dorsales (voir section 1.3.5). Ces activations ont été confirmées pour d'autres types de modifications d'images, au niveau de la texture, en ajoutant du bruit, ou en détruisant l'information de phase dans le domaine de Fourier. Le LOC est donc une région centrale montrant une activation spécifique aux objets familiers et non-familiers (Grill-Spector *et al.*, 2001). L'aire VOT pourrait être une région encore plus spécialisée, on y trouve par exemple deux régions bien spécifiques : une région plutôt fovéale et sélective aux visages, la FFA (Fusiform Face Area) (Kanwisher *et al.*, 1997) et une région plutôt périphérique et associée aux lieux, la PPA (Parahippocampal Place Area).

L'aire LOC correspondrait donc à l'aire la plus avancée hiérarchiquement, encore située dans le cortex occipital. L'information visuelle serait ensuite diffusée dans des régions qui vont progressivement devenir de plus en plus associées aux aspects mnésiques (la fin du chemin se trouvant dans l'hippocampe). L'aire VOT décrite par Grill-Spector, se situe dans la partie ventrale du cortex temporal. Ce qui correspond à une région très largement étudiée, tout particulièrement chez le singe : le cortex inféro-temporal (IT). Des enregistrements pionniers ont permis de montrer que les neurones d'IT répondaient de façon sélective à des stimuli complexes comme des visages (Perret *et al.*, 1982; Gross *et al.*, 1972; Desimone *et al.*, 1984). Les expériences du groupe de Logothetis sur l'apprentissage de nouveaux objets suggèrent que cette représentation serait plastique et que les neurones de IT pourraient apprendre à répondre à de nouvelles classes de stimuli (Logothetis et Sheinberg, 1996). L'équipe de Tanaka a, quant à elle, montré que ces neurones peuvent aussi répondre à une multitude d'objets différents,

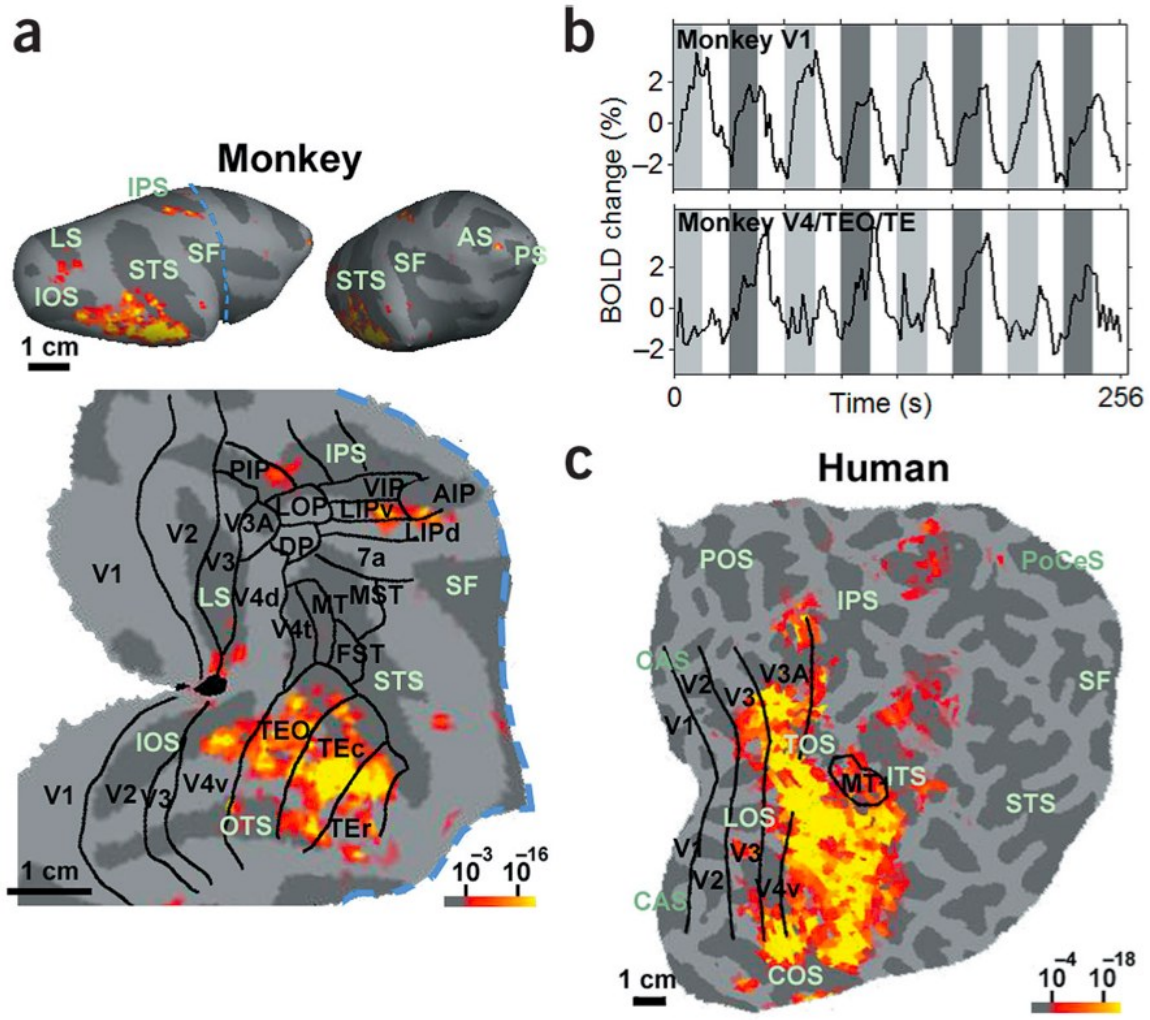

FIGURE 1.4: Activation des aires sélectives aux objets chez le macaque et l'humain. **a)** Aires chez le macaque significativement plus activée par la présentation d'objets intacts que par des images où l'on mélange les parties de l'objet (anglais : grid-scrambled). Ces activations significatives arrivent principalement dans les aires V4, TEO et TE. **b)** Décours temporal de l'activation de V1 et des aires V4/TEO/TE lors de la présentation alternée d'objets intacts (gris foncé), mélangés (gris clair) ou d'un écran vide (blanc). On voit clairement que l'aire V1 s'active dès lors qu'un image est présentée, alors que les aires V4/TEO/TE s'activent sélectivement pour l'objet intact. **c)** Activités dans l'hémisphère droit d'un sujet humain avec le même type de comparaison. Extrait de Tsao *et al.* (2003).

qu'ils seraient organisé de façon modulaire. À l'intérieur de IT, les représentations des objets deviennent de plus en plus abstraites (Tanaka, 1996). Il est ainsi inapproprié de considérer IT comme une structure unique. On peut en effet clairement le diviser en deux parties bien distinctes : sa partie postérieur (PIT, ou TEO chez le singe) et sa partie antérieur (AIT, ou TE chez le singe). La partie postérieure (PIT) serait une étape intermédiaire avant la partie antérieure (AIT). On aurait ainsi une représentation plus complexe dans PIT que dans V4 mais pas encore vraiment invariante à la position et à la taille. Les neurones d'AIT sont généralement considérés comme représentant des conjonctions complexes de caractéristiques, comme le montre des études

lésionnelles (Moss *et al.*, 2005) ou d'IRMf (Tyler *et al.*, 2004) et seraient particulièrement importants pour la discrimination entre les objets au niveau individuel. De plus, il semble que différentes parties de AIT coderaient différents objets selon une hiérarchie assez intuitive. D'abord les objets animés sont dans une région bien séparée des objets non-animés. Ensuite, parmi les objets animés, les visages, le mains et les parties du corps seront dans des régions différentes. Parmi les visages, les visages primates et les visages non-primates sont aussi séparés. Le groupe des visages primates serait aussi séparé entre humain et singes. Il semble ainsi que l'organisation des neurones de AIT recoupe largement notre structure intuitive des catégories (Kiani *et al.*, 2007).

Cependant, l'organisation rétinotopique n'a plus lieu dans AIT, ce qui permet en contrepartie une invariance à la position. La réponse des neurones d'IT est ainsi relativement invariante à la position dans le champ visuel, ainsi qu'à la taille et la vue (Gross *et al.*, 1972), résultant en des champs récepteurs extrêmement larges en comparaison avec les aires rétinotopiques en amont (jusqu'à 50° d'angle visuel). Le niveau de complexité et d'invariance atteint par les neurones d'AIT suggère qu'ils seraient des candidats sérieux pour le rôle de neurone "grand-mère" (un neurone = un concept). Pour étudier cette question, Hung et al. ont utilisé une technique appelée décodage, basé sur l'entrainement puis le test d'un classifieur mathématique sur des données neurophysiologiques. Ils ont ainsi enregistré l'activité ("spiking activity") de groupes de neurones dans AIT pendant que les singes regardaient des images de différents objets. Ils entrainaient ensuite le classifieur sur une partie des données, puis testaient ses performances pour extraire différents types d'information sur le reste des données. Il ont ainsi pu montrer que l'on pouvait, à partir de l'activité des neurones de AIT, déterminer la catégorie de l'objet présenté de façon précise et même l'identifier parmi les autres exemplaires de ce groupe (Hung *et al.*, 2005). L'information serait donc contenue dans le patron d'activation des neurones de AIT, et non pas dans un seul neurone. Ce codage par population constitue une base efficace pour la reconnaissance de l'objet unique.

Le cortex inféro-temporal est la dernière région associée exclusivement aux traitements visuels. Il se trouve déjà dans le cortex temporal, où siège un grand nombre d'aires largement associés aux processus mnésiques. L'information va ensuite être traitée par des aires multi-modales qui vont aussi être largement activée par d'autres sens, mais surtout être impliquées dans la mémoire : les structures du lobe temporal médian.

1.3.3 Au bout du chemin : le lobe temporal médian

Comme on peut le voir dans la Figure 1.5, toutes les régions de la voie visuelle se projettent finalement vers les différentes structures composant le lobe temporal médian (MTL) : *amygdale*, *cortex enthorinal*, *cortex parahippocampique*, *cortex perirhinal* et enfin *hippocampe*. L'hippocampe reçoit des afférences du cortex entorhinal, qui lui-même

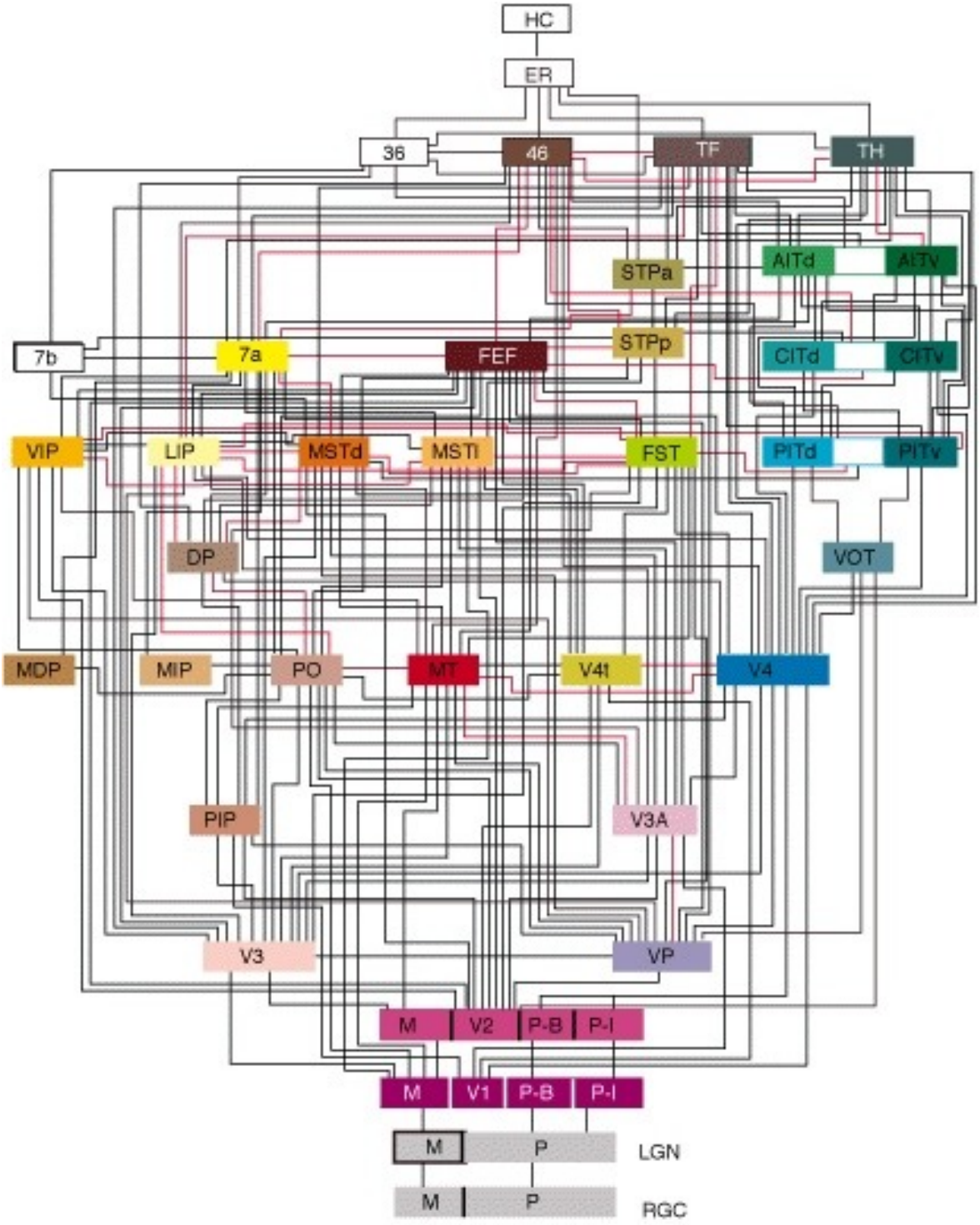

FIGURE 1.5: L'organisation du système visuel chez le primate, selon Felleman et Van Essen (1991).

reçoit la majorité de ses afférences du cortex perirhinal, du cortex parahippocampal et dans une moindre mesure directement de IT. Ces régions ne sont pas purement visuelles et leur rôle a souvent été associé à la consolidation en mémoire à long-terme des informations en mémoire à court-terme (Squire *et al.*, 2004). Cependant, une série d'études récentes a été menée pour étudier leur rôle dans les traitements visuels. Ceci

a été possible chez l'humain grâce à l'implantation de micro-électrodes chez des sujets épileptiques en attente d'une opération. Il a ainsi été montré qu'un grand nombre de cellules de ces régions répondaient sélectivement à différentes catégories d'objets comme des visages, des animaux ou des maisons (18% de l'hippocampe, 16% du cortex enthorinal et 9% de l'amygdale) (Kreiman *et al.*, 2000). Une autre étude a été plus loin en montrant qu'une grande partie de ces neurones (40%) répondaient de façon extrêmement sélective et invariante à des objets précis. On peut voir l'exemple dans la Figure 1.6a d'un neurone qui répond non seulement aux photos de Halle Berry, mais aussi à l'écriture de son nom, tout en ne répondant pas aux photos d'autres personnes (Quiroga *et al.*, 2005).

Cette même méthodologie a ensuite été utilisée pour savoir si l'activation de ces neurones pouvait être assimilé à ce que Christof Koch appelle les *corrélats neuronaux de la conscience* (Koch, 2004). Autrement dit, est ce que l'activité de ces neurones va être corrélée directement avec la perception consciente des sujets. Ils ont pour ceci utilisés un protocole de suppression par flash, dérivé de la rivalité binoculaire, permettant de changer la perception consciente sans changer la stimulation au niveau rétinien. Ils ont ainsi montré que l'activité de 2/3 des neurones qu'ils ont enregistrés était corrélée avec les changements de perception du sujet, plutôt qu'avec les changements au niveau rétinien. Les neurones sélectifs ne répondaient pas lorsque le stimulus était présent mais pas perçu. Il semble donc que l'activité des neurones du lobe temporal médian est directement corrélée à l'expérience visuelle phénoménale des sujets (Kreiman *et al.*, 2002).

Ces observations tendraient à considérer les neurones de cette région comme de parfaits candidats pour le rôle de *neurone grand-mère*. L'hypothèse du codage de la représentation des objets sous la forme d'un neurone grand-mère est la version extrême de l'hypothèse de *codage éparse*, par opposition avec un *codage distribué*.

Codage distribué : Un percept donné est codé par l'activité d'un grand ensemble de neurones, dans lequel chaque neurone est sélectif à une caractéristique particulière. Dans ce cas, pour chaque objet, une grand nombre de neurone de la population vont participer à son codage.

Codage éparse : Dans ce cas, un même percept est codé par un nombre beaucoup plus restreint de neurones. Dans ce cadre, la majorité des neurones ne répondent pas à la majorité des objets. La version extrême étant l'hypothèse du neurone grand-mère (1 neurone = 1 concept).

Bien que les cellules présentées précédemment semblent à première vue se comporter comme des neurones grand-mères, plusieurs arguments vont plutôt dans le sens d'un codage plutôt de type éparse des cellules du MLT (Quiroga *et al.*, 2008). Premièrement, s'il n'y avait vraiment qu'une cellule codant pour Halle Berry par exemple, les expérimentateurs auraient eu une chance incroyable de la trouver. Ensuite, il n'est pas

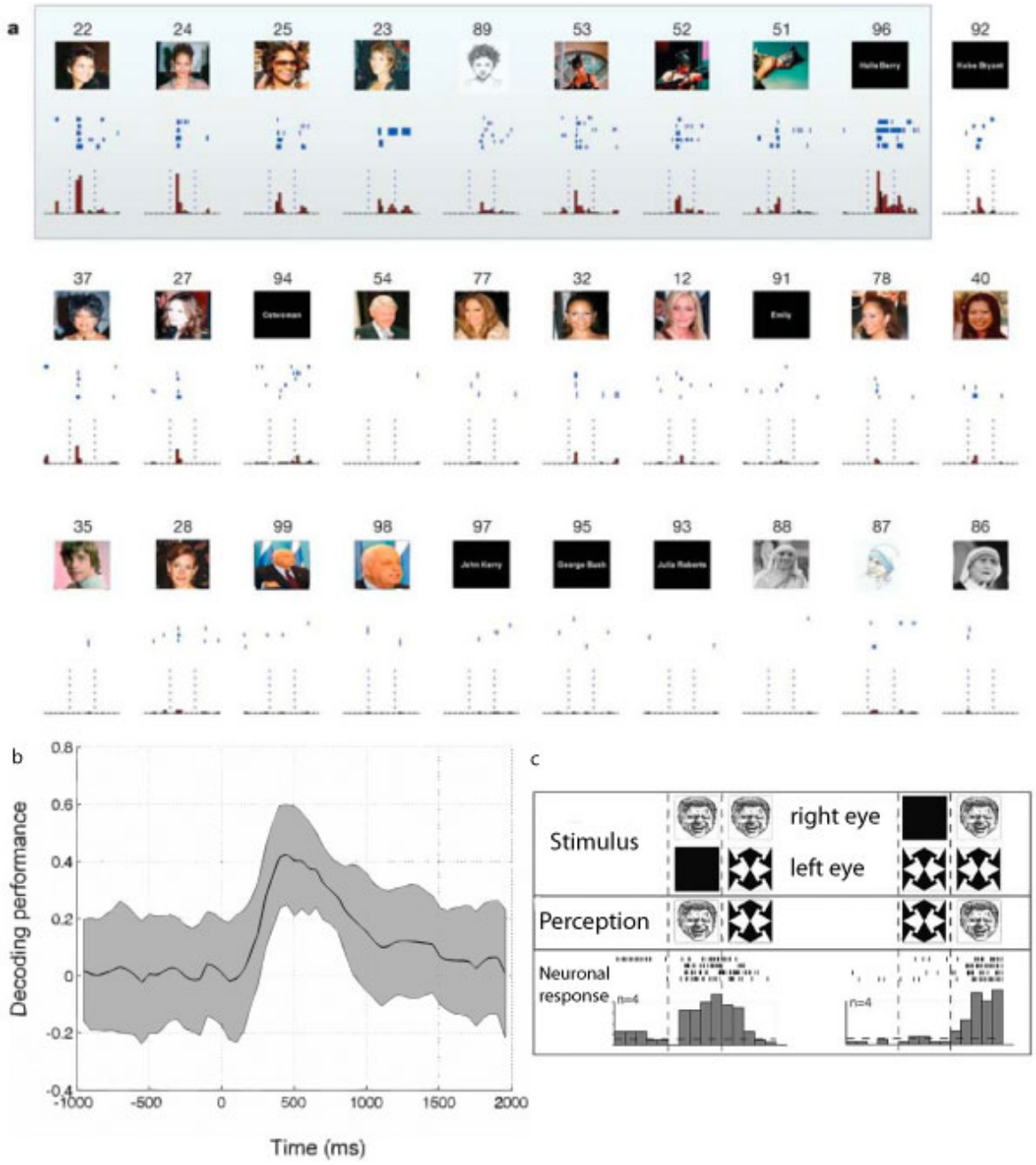

FIGURE 1.6: **Les neurones du lobe temporal médian. a)** Exemple de cellule de l'hippocampe postérieur gauche qui répond sélectivement et exclusivement aux images de Halle Berry (Quiroga *et al.*, 2005). **b)**Décours temporel des performances d'un classifieur pour décoder l'information sur la catégorie de l'objet à partir de l'activité des neurones du lobe temporal médian (Quiroga *et al.*, 2007). **c)** Protocole de suppression par flash, et activité d'un neurone sélectif à Bill Clinton selon les différentes conditions (Kreiman *et al.*, 2002). On voit que le neurone s'active en parfaite correspondance avec la perception consciente du sujet, et non avec ce qui est présenté à la rétine.

certain que cette même cellule ne répondrait pas pour d'autres personnes. Les temps d'expérimentation dans ces expériences étant assez court, et le nombre d'images présentées limité, il est n'est pas possible d'exclure que certaines de ces cellules auraient répondues à d'autres stimuli. Les auteurs ont ainsi trouvé une cellule répondant sé-

lectivement à Jennifer Aniston, mais aussi à Lisa Kudrow (deux actrices de la série télévisée Friends) ainsi qu'une autre cellule répondant à la Tour Eiffel et à la Tour de Pise (Quiroga *et al.*, 2005). Il est intéressant de noter que ces associations ne sont pas sans sens.

Ces arguments suggèrent donc l'existence d'un codage très éparse de l'information dans les neurones du MTL, ce qui a aussi été appuyé par une étude théorique (Waydo *et al.*, 2006). Ce même groupe a ensuite utilisé des outils mathématiques de décodage pour "deviner" l'identité de l'image à partir de l'activité de l'ensemble des cellules de cette région. Leurs résultats montrent que l'information augmente avec le nombre de cellules prises en compte, et surtout qu'elle est optimale entre 400 et 500 ms (Quiroga *et al.*, 2007). Il est important de remarquer que les premières activations des neurones du MLT sont extrêmement tardives en comparaison aux activations des neurones de IT (voir section 1.3.6). Ce qui se passe durant cet intervalle de temps reste donc à définir.

1.3.4 Modèle feedforward de la voie visuelle ventrale

L'approche computationnelle en sciences de la vision a été largement initiée par le travail de David Marr, qui en a posé les fondements, et reste aujourd'hui un auteur majeur du domaine (Marr, 1982). Son approche a consisté à théoriser le fonctionnement de différents système cérébraux, notamment le système visuel, à l'aide d'outils mathématiques. À l'heure actuelle, la neurophysiologie des différentes aires de la voie ventrale, ainsi que l'informatique en général, ayant fait des avancées considérables, cette approche est toujours d'actualité. Il est donc possible maintenant de créer des modèles informatiques très précis et directement basés sur l'anatomie et la physiologie des neurones de la voie visuelle. Ceci permet aussi en retour de tester la validité des hypothèses émises à partir des données neurophysiologiques.

Le but d'un modèle de reconnaissance d'objets à l'heure actuelle serait donc de simuler l'interaction entre les différentes aires de la voie ventrale. Le groupe de Poggio a travaillé dans ce sens. Avec une approche purement feedforward (information ascendante dans la voie visuelle d'un point de vue hiérarchique) censée expliquer la toute première vague d'information dans la voie ventrale. Il n'y a donc ici aucun processus feedback (information descendante). Ce modèle se base sur quatre propriétés principales (Serre *et al.*, 2007a) :

- Une construction hiérarchique de l'invariance, d'abord à la position, puis à la taille, et enfin au point de vue.
- Une augmentation de la taille des champs récepteurs qui s'accompagne d'un augmentation de la complexité de leur structure.
- Un traitement feedforward basique, pas de feedback.
- Plasticité et apprentissage prennent place à tous les niveaux, à une échelle de

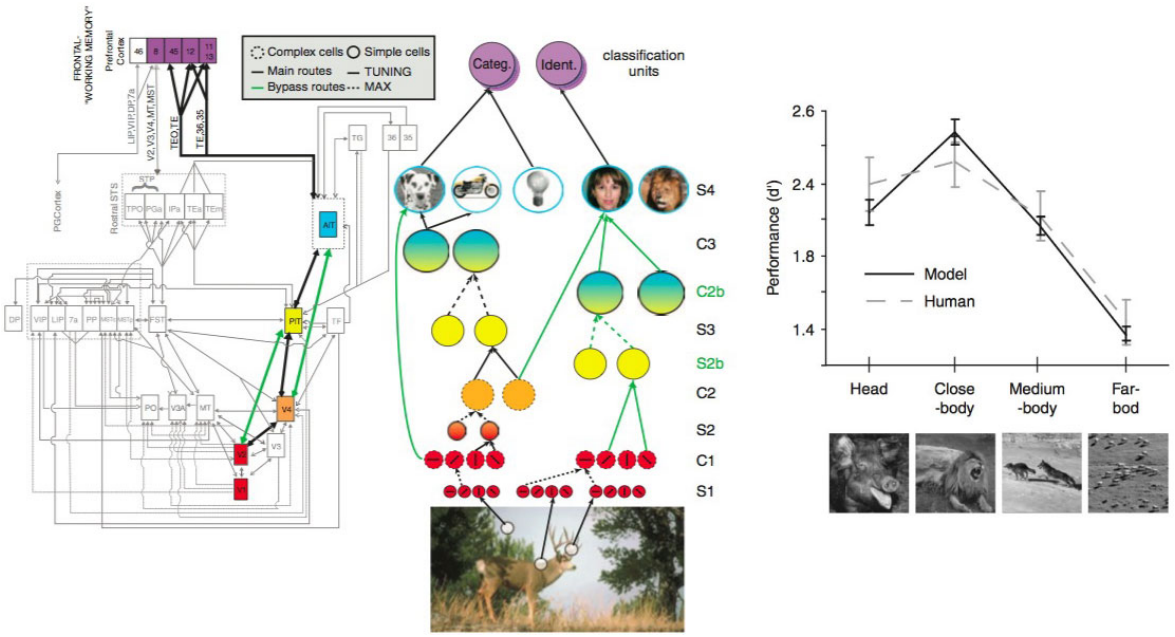

FIGURE 1.7: Le modèle de Serre & Poggio **Sur la gauche :** Les différentes couches du modèle et leur correspondance avec les aires cérébrales. **Sur la droite :** Performances du modèle et de sujets humains dans une tâche de catégorisation animal/non-animal en fonction de la taille occupée par l'objet dans l'image. Comme on peut le voir, le modèle présente le même patron de performance. Extrait de Serre *et al.* (2007a).

temps qui décroit entre V1 et IT.

Le but du modèle étant d'avoir la meilleure balance possible entre invariance et sélectivité, il se base sur deux types de couches différentes qui auront chacune un rôle particulier.

Les unités S simples ont une sélectivité en forme de gaussienne, caractéristique répandue dans tout le cortex visuel. En effet, la sélectivité à une orientation précise des neurones de V1 est une simplification. Si l'on regarde précisément leurs réponses en fonction de l'orientation (leurs courbes de sélectivité), celles-ci ont une forme de gaussienne autours de leur orientation préférée. Ceci s'observe aussi pour les neurones de AIT, qui sont sélectifs à un point de vue particulier du visage, mais montrent aussi une sélectivité en forme de gaussienne pour les autres vues de cet objet (Perrett *et al.*, 1998). Dans le modèle, un neurone S reçoit donc ses afférences de cellules ayant différentes sélectivités pour une même position dans le champ visuel, et les assemble de façon linéaire. Cette couche permet donc de gagner en sélectivité et en complexité au fur et à mesure que l'information avance dans la voie.

Les cellules C complexes vont opérer une fonction non-linéaire MAX entre leurs afférences (Riesenhuber et Poggio, 1999; Rousselet *et al.*, 2003b). Une cellule de la couche C recevra ainsi ses afférences d'un ensemble de cellules S similaires mais différant sur un paramètre (par exemple la taille) auquel la cellule C deviendra donc invariante. Les caractéristiques des couches C permettent donc l'invariance à un ensemble de paramètres

(taille, position, point de vue).

Une bonne façon de tester la pertinence d'un modèle informatique d'une fonction cognitive est de comparer ses variations de performances avec celles des humains. Les auteurs ont donc comparé les performances du modèle et des sujets dans une même tâche de catégorisation animal/non-animal. Plus précisément, ils ont regardé comment ces performances variaient en fonction de la taille occupée par l'animal dans la scène. Comme on peut le voir sur la figure 1.7, les sujets humains et le modèle montrent exactement le même patron de performance (Serre *et al.*, 2007b).

1.3.5 Un rôle pour certaines aires pariétales ?

Bien que les traitements associés à la reconnaissance d'objets soient largement associés à la voie ventrale, des activations sélectives aux objets ont aussi été trouvées dans des aires de la voie dorsale. Ainsi, dans l'étude en IRMf de Grill-Spector, recherchant les aires spécifiques aux objets dans le cortex humain, la majorité étaient localisées dans les cortex occipitaux et temporaux. Cependant, certaines aires de la voie dorsale se trouvaient aussi activées : notamment une région proche de V3a, et une autre antérieur à V3a et recouvrant partiellement V7 (Grill-Spector, 2003).

De plus cette question rejoint une autre question d'ordre plus général. Comme nous l'avons vu, la sélectivité et l'invariance atteintes par la voie ventrale dans la reconnaissance d'objets a un coût qui se manifeste par la grande taille des champs récepteurs dans les aires supérieures comme IT. Cette reconnaissance se fait donc au prix de la localisation, un système basé uniquement sur la première vague d'activation de la voie ventrale de V1 à IT peut certainement être très bon pour reconnaître les objets, mais il perdra alors l'information sur sa localisation précise. Ce rôle de localisation est généralement attribué à la voie dorsale du système visuel. Une interaction entre ces deux systèmes pourrait donc être nécessaire pour aboutir à une perception localisée et complète des objets.

Les neurones des aires visuo-motrices comme LIP (cortex Latéral IntraPariétal) et FEF (Frontal Eye Field, qui n'est donc pas dans le cortex pariétal mais dans le frontal) montrent souvent un codage sélectif aux différentes cibles dans le champ visuel (voir section 1.6). Cependant, ce codage est le plus souvent associé à la tâche en elle-même, ces neurones deviendraient ainsi des détecteurs provisoires de caractéristiques en fonction de la pertinence des stimuli pour la tâche à réaliser. Ainsi, lors d'une tâche où les sujets doivent classer différents mouvements en deux catégories, les neurones de MT (l'aire sensorielle où l'on trouve des neurones sélectifs aux différentes directions de mouvement) codent toutes les possibilités de mouvement, alors que les neurones de LIP vont s'activer en classant les mouvements en deux catégories (Freedman et Assad, 2006). Ceci ne correspond donc pas à un vrai codage des caractéristiques des stimuli

mais plutôt à un codage provisoire de leur pertinence pour la tâche à effectuer.

Un certain nombre d'études semblent cependant avoir trouvé un codage de la forme au sens propre dans ces aires visuo-motrices. En utilisant un protocole similaire à ce qui se fait pour la reconnaissance d'objets, il a ainsi été montré que des neurones de LIP (Sereno et Maunsell, 1998; Janssen *et al.*, 2008) et de FEF (Bichot *et al.*, 1996; Peng *et al.*, 2008; Kirchner *et al.*, 2009) répondaient sélectivement à des formes simples (lettres, formes géométriques comme des carrés et des ronds). Au niveau anatomique, de nombreuses connections existent entre les aires purement visuelles et ces aires visuo-motrices. Par exemple, LIP reçoit de nombreuses projections des aires V4 et IT (Webster *et al.*, 1994). Il semble donc que l'on puisse avoir un codage basique de la forme dans la voie dorsale et dans les aires visuo-motrices. Comme nous le verrons par la suite, ceci pourrait servir à une reconnaissance et une localisation rapide d'objets dans le champ visuel, spécialement car ces aires s'activent très précocement après la présentation de stimuli (voir section 1.3.6).

1.3.6 Timing des réponses cérébrales et feedback

L'aspect temporel est une donnée importante pour la compréhension des mécanismes de perception visuelle. Récemment, une étude a décodé l'activité de IT chez le singe en fonction du temps (Hung *et al.*, 2005). Ils montrent de cette façon que la catégorie de l'objet ainsi que son identité peuvent être lues à partir de l'activité des seuls neurones d'IT de façon fiable. Plus précisément, il ressort que la catégorie de l'objet peut être lue dans IT seulement 100 ms après la présentation de l'image en ne prenant en compte que l'activité sur une fenêtre temporelle de 12,5 ms. Chez l'homme, une étude du même type a aussi pu être menée chez des patients épileptiques. Chez des patients implantés avec des électrodes proches du cortex inférotemporal, les auteurs ont analysé la réponse des neurones de cette aire lors de la présentation de différentes catégories d'images. Ils ont aussi pu montrer que même chez l'homme, l'activité des neurones de IT contenait de l'information sur la catégorie d'objet dès 100 ms (Liu *et al.*, 2009). Il est cependant important de rappeler que ce n'est pas parce que l'on peut décoder une information dans l'activité d'une aire que cette information est déjà exploitée par le système. En effet, les enregistrements de potentiels locaux sont largement influencés par les variations électriques au niveau synaptique. L'information n'est donc pas encore transmise par les neurones mais encore en train d'être intégrée. Il faut attendre les premiers potentiels d'action pour vraiment considérer l'information comme disponible pour les aires suivantes. Il faut donc prendre des précautions avec ces résultats, car il est probable que l'on puisse "décoder" une information bien avant qu'elle soit effectivement disponible pour les aires suivantes (une différence qui pourrait être de l'ordre de 20/30 ms, Monosov *et al.*, 2008).

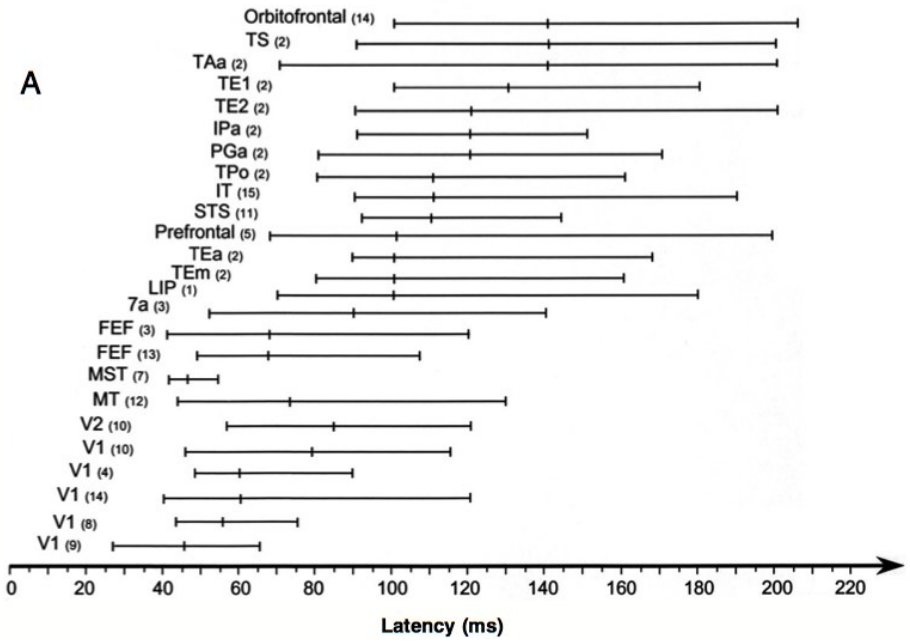

FIGURE 1.8: Les latences d'activation des différentes aires du système visuel chez le singe. Extrait de Bullier (2004).

L'aspect temporel peut aussi servir à mieux comprendre les relations de causalité entre les différentes aires du système visuel. Si la hiérarchie que nous avons présentée jusqu'ici est correcte, alors la rétine va s'activer en premier, puis le LGN, V1, V4 et enfin IT avant d'avoir des réponses dans le lobe médian temporal. Si l'information est transmise de façon séquentielle d'aire en aire, alors ceci doit être traduit par un ordre d'activation cohérent. Il est cependant très complexe au niveau expérimental de tester dans une même étude les latences des neurones de ces différentes aires lors d'une même stimulation. Cependant plusieurs méta-études ont été menées afin de synthétiser les temps de réponses moyens, ainsi que les toutes premières activations des différentes aires de la voie visuelle (Bullier, 2004; Lamme et Roelfsema, 2000). Comme on peut le voir dans la figure 1.8, l'ordre d'activation respecte globalement la hiérarchie des aires abordées auparavant. Cependant, on peut aussi noter que certaines aires s'activent beaucoup plus précocement que ce que leur place dans la hiérarchie laisserait penser, notamment MT (aire sensorielle associée à la perception du mouvement) ainsi que FEF et LIP. Il semble donc que ces aires puissent recevoir de l'information très précocement (certainement via des afférences sous-corticales), mais plus important, elle peuvent donc très tôt envoyer de l'information en retour vers les aires inférieures. Il se peut donc qu'un activité feedback arrive très tôt et puisse même venir moduler la première vague d'activation de la voie ventrale. L'équipe de Jean Bullier a largement étudié le rôle possible de ce feedback précoce. Selon leur théorie, l'information rétro-injectée en feedback pourrait avoir plusieurs formes :

– L'information globale modulerait l'activité locale des neurones des aires précoces

(V1, V2, V3)
- Les différentes voies n'ayant pas toute la même vitesse de conduction (magnocellulaire plus rapide que parvocellulaire, lui-même plus rapide que koniocellulaire), l'information contenue dans les voies les plus rapides servirait à moduler l'activité des suivantes.
- En allant encore plus loin, différents types d'indices bas-niveau seraient traités à différentes vitesses (le mouvement avant la forme, la couleur encore plus tard). Si la tâche peut-être basée sur plusieurs de ces indices (reconnaître une boule rouge sur un fond vert). L'information de l'un pourrait servir de "guide" aux autres en fonction de leur date d'arrivée.

L'hypothèse d'un rôle fondamental des connexions feedback a notamment été mise en avant à partir de modèles computationels du traitement cérébral, autant pour les connexions thalamo-corticales (Mumford, 1991) que cortico-corticales (Mumford, 1992). Il a aussi été suggéré que l'aire V1 pourrait être bien plus qu'une porte d'entrée de l'information dans le cortex. L'activité des neurones de V1 pourrait ainsi être modulée plutôt précocement par la ségrégation figure/fond (Hupé *et al.*, 1998; Roelfsema *et al.*, 2007) ou pour la perception de la forme par l'ombre (Lee *et al.*, 2002). Les boucles récurrentes entre V1 et les aires supérieures serviraient ainsi à intégrer les *a priori* contextuels descendants à l'information ascendante (Lee et Mumford, 2003).

Concernant la reconnaissance d'objets, le groupe de Nancy Kanwisher s'est intéressé récemment à ces phénomènes. Ils ont d'abord montré en IRMf (il semble que ce soit à la base une découverte fortuite) que l'information sur un objet présenté en périphérie pouvait être retrouvée dans l'activité des aires visuelles rétinotopiques fovéales (Williams *et al.*, 2008). Une autre étude récente a étudié la perception des images de "Mooney" en IRM fonctionnelle (Hsieh *et al.*, 2010). Ces images composées uniquement de pixels noirs ou blancs (et pas en niveau de gris) ne permettent souvent pas de reconnaître l'objet à première vue. Cependant, une fois que l'on a vu l'image originale, on reconnait très bien l'image de "Mooney". Les auteurs ont ainsi montré que l'activité dans V1 pour le "Mooney" reconnu était plus semblable à l'activité pour l'image originale que pour le même "Mooney" non reconnu. Ceci est une preuve évidente que l'activité de V1 n'est pas une représentation du champ visuel uniquement conduite par les traits physiques de l'image, mais est influencée par de nombreux processus descendants.

1.4 Dynamique de la perception

Comme nous venons de le voir, l'aspect temporel est primordial dans l'approche de la perception. Malheureusement, ce facteur n'est souvent pas assez mis en avant. L'approche classique de la psychophysique est, à l'origine, de comprendre le fonctionnement

des systèmes sensoriels en testant leurs limites. Pour ceci, les protocoles historiques de perception au seuil et de "différence juste notable" ont été mis en place pour tester les limites du système tant au niveau spatial qu'au niveau de la sensibilité aux paramètres de luminance et de contraste. Les études s'intéressant aux limites temporelles ont été beaucoup moins nombreuses. Nous allons dans cette section nous y intéresser de plus près.

1.4.1 Catégorisation rapide de scènes naturelles

Une des particularités de notre système visuel est qu'il fonctionne extrêmement rapidement. Ainsi, la perception de notre monde environnant se fait de façon immédiate et sans effort. La première à s'être penchée explicitement sur cette dynamique pour la perception de scènes naturelles a été Marie C. Potter à la fin des années 60 (Potter et Levy, 1969), créant par la même occasion un protocole qui sera largement utilisé par la suite. L'expérience consistait à présenter aux sujets une série d'images différentes à un rythme plus ou moins rapide (Rapid Serial Visual Presentation, RSVP). Ceci permettait ensuite de déterminer la "résolution temporelle" du système visuel pour différentes tâches en fonction de la fréquence de présentation des images : détecter la présence d'un objet cible dans la série, dire si une image présentée a posteriori faisait partie de la série. Ces expériences ont ainsi permis de mettre en évidence de façon quantitative les performances temporelles du système visuel.

Une autre étude importante, qui a notamment permis de poser une contrainte claire sur les modèles de reconnaissance d'objets, est celle de Thorpe, Fize & Marlot en 1996 (Thorpe *et al.*, 1996). Cette étude a véritablement lancé cette thématique dans l'équipe de recherche dont je fais partie, tout en stimulant largement le domaine en général. Dans cette expérience, à chaque essai, une image était présentée durant un temps très court de 20 ms, pour éviter toute exploration oculaire. Le sujet devait alors simplement signaler le plus vite possible, en relevant le doigt d'un détecteur infra-rouge, si la photo contenait un animal (*tâche go/no-go*). La moitié des images contenait effectivement un animal, l'autre moitié était constituée d'images extrêmement variées (paysages, véhicules, panneaux de signalisation, monuments, etc). Ils ont ainsi pu montrer que les premières réponses "fiables" apparaissaient avant 300 ms (i.e. le premier intervalle de temps où les réponses correctes étaient significativement supérieures aux réponses incorrectes, ce que l'on appellera le temps de réaction minimum). En enregistrant l'activité EEG (électro-encephalographie), ils ont montré qu'une activité différentielle entre les essais contenant une cible et ceux contenant un distracteur apparaissait dès 150 ms lors d'une analyse en potentiels évoqués (ERP). Vu sa latence, cette activité différentielle était un support parfait pour les réponses manuelles rapides dès 250 ms.

Ces temps posaient de nouvelles contraintes importantes à plusieurs niveaux dans

le domaine de la reconnaissance d'objets. Le point de vue classique voudrait que la reconnaissance d'objets dans une scène complexe nécessite au préalable de traiter la scène dans son ensemble, puis d'isoler les différents objets (ségrégation figure-fond) pour enfin comparer les objets isolés au modèle de l'objet que l'on recherche. Il est très difficile d'imaginer que ces nombreuses opérations soient réalisées en seulement 150 ms. En effet, les contraintes physiologiques doivent aussi être prises en compte. Si l'on considère ainsi que l'information a besoin de 10 ms pour traverser une couche, et qu'elle doit traverser environ 10 couches (ou synapses) entre la rétine et les aires impliquées dans la reconnaissance d'objets comme le cortex inférotemporal (IT), on arrive alors vite à la conclusion qu'il y a très peu de temps à laisser en chemin (Thorpe et Fabre-Thorpe, 2001). Il est alors très compliqué d'imaginer la possibilité de boucles pour ces réponses rapides, l'information doit aller droit devant (traitement purement feedforward). Cette contrainte temporelle forte suggère donc qu'une seule et unique vague d'information est suffisante pour traiter et analyser un scène aussi complexe qu'une scène naturelle et détecter la présence d'un animal dans cette scène.

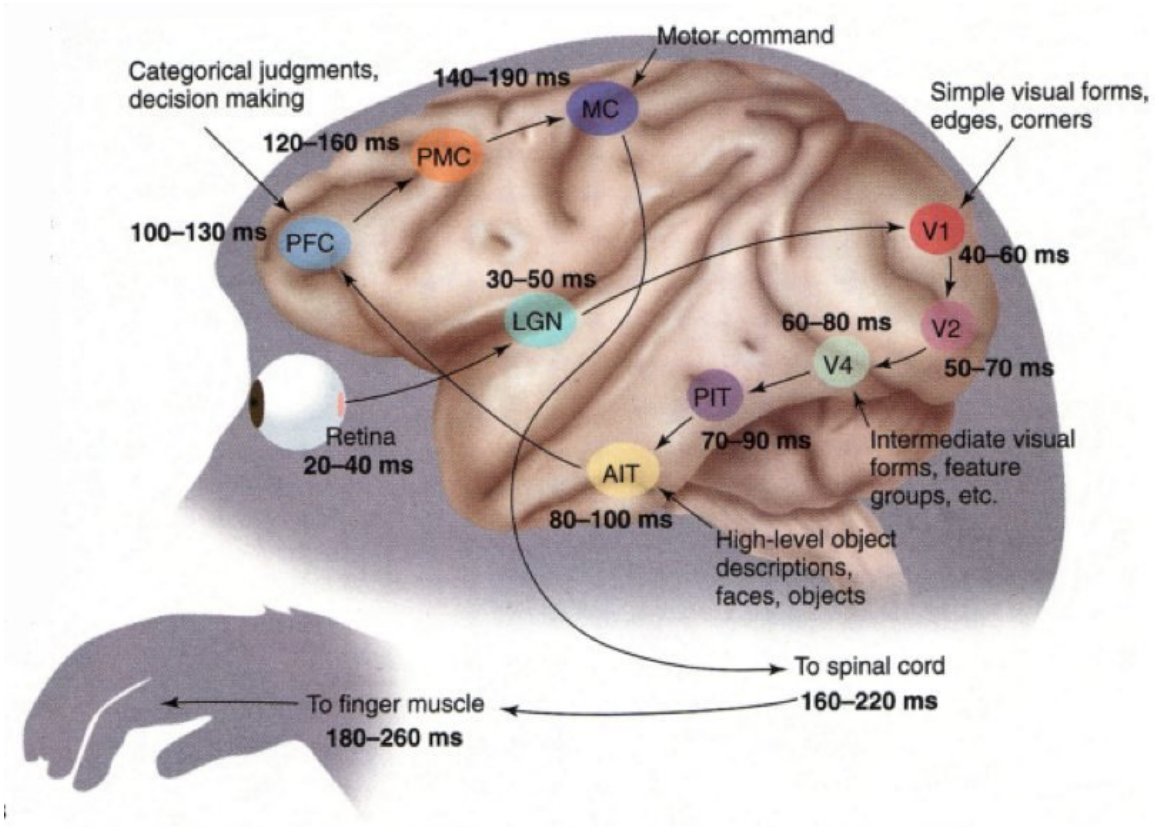

FIGURE 1.9: Les différentes étapes que doit parcourir l'information visuelle depuis la rétine jusqu'à la réponse motrice. Les latences à côté de chaque aire sont celles des toutes premières réponses de cette aire, et leur valeur moyenne. Extrait de Thorpe et Fabre-Thorpe (2001).

Ces résultats tendaient à supporter une autre hypothèse forte, celle du codage temporel par rang (Thorpe, 1990). En effet, le code neuronal est le plus souvent considéré comme contenu dans le taux de décharge, qui serait le moyen utilisé par les neurones

pour communiquer entre eux. Cependant, encore une fois, des temps de réponse si courts supposent une transmission extrêmement rapide de l'information entre chaque couche. Le problème du taux de décharge est qu'il repose sur une mesure de fréquence et nécessite donc une fenêtre temporelle d'intégration, ce qui prendrait beaucoup de temps à chaque étape. Il a donc été proposé qu'un autre mode de communication puisse être utilisé par les neurones : le codage par rang (Gautrais et Thorpe, 1998). Selon celui-ci, de l'information est aussi contenue dans l'ordre d'arrivée des potentiels d'action afférents à un neurone. Les neurones les plus activés (i.e. les zones les plus "saillantes" de la scènes) seraient les premiers à décharger. Cette transmission de l'information rapide et efficace serait un support plausible aux temps de réactions très courts observés dans ces tâches de catégorisation (VanRullen et Thorpe, 2002; Vanrullen *et al.*, 2005).

Par la suite, un grand nombre d'études ont été réalisées pour mieux comprendre ces mécanismes de catégorisation rapide. Il a ainsi été montré qu'ils ne pouvaient pas être accélérés, même après un apprentissage intensif des images (Fabre-Thorpe *et al.*, 2001). Des sujets ont ainsi quotidiennement, pendant trois semaines, passé une expérience de catégorisation rapide avec les mêmes images chaque jour. Lorsque à la fin de ces trois semaines, ils ont été testés sur ces mêmes images mélangées à des images nouvelles, leurs temps de réaction moyens étaient plus court sur les images apprises (notamment par la diminution des réponses les plus lentes) mais leurs temps minimum, ainsi que l'activité différentielle à 150 ms sont restés inchangés. Le processus de traitement rapide serait donc déjà optimal, ce qui est un nouvel argument en faveur d'une vague purement feedforward. Un autre argument est venu d'une expérience utilisant un protocole similaire mais où les images étaient "masquées". Le masquage rétrograde (le seul qui sera abordé ici) consiste à afficher à la suite de l'image cible une autre image (en l'occurrence ici des images artificielles à fort contraste et à différentes fréquences spatiales) qui va perturber son traitement. En général, on considère que le masquage rétrograde bloque les boucles feedback prenant plus de temps que l'intervalle entre l'image cible et le masque. Avec ce protocole, Bacon-Macé et al. ont montré que les sujets n'étaient plus capable de détecter la présence d'un animal lorsque le masque suivait immédiatement l'image cible, mais qu'avec un intervalle de 40 ms, ils atteignaient un niveau tout à fait respectable de 75% correct. La capacité de détection résiste donc fortement à la suppression supposée des boucles feedback (Bacon-Macé *et al.*, 2005). Le temps de traitement nécessaire pour catégoriser l'image comme contenant un animal ne serait même pas plus long que pour simplement détecter la présence d'un objet non défini dans l'image (Grill-Spector et Kanwisher, 2005). Cette équivalence entre reconnaissance et détection est cependant à relativiser car si on rend la tâche plus difficile (par exemple avec du bruit), la reconnaissance est ralentie, contrairement à la détection (Mack *et al.*, 2008).

Un autre aspect de ces traitements est qu'ils opèreraient de façon massivement

parallèle. Dans une expérience similaire aux précédentes, où les sujets devaient lever le doigt s'ils voyaient un animal, Rousselet et al. ont multiplié le nombre de scènes à traiter simultanément et ont montré que ceci pouvait être fait sans aucune chute de performance lorsque deux images étaient présentées en même temps à l'écran (Rousselet *et al.*, 2002, 2004b,a). Le performance des sujets diminuait ensuite significativement avec l'augmentation du nombre d'images. Ces résultats suggèrent un traitement global et parallèle du champ visuel. La catégorisation rapide ne nécessiterait pas non plus d'information chromatique (Delorme *et al.* 2000 ; même si selon Gegenfurtner et Rieger 2000, elle améliorerait la mémorisation), fonctionneraient très bien à bas contraste (Macé *et al.*, 2005), et pourraient être réalisée même lorsque les images sont présentées avec une très grande excentricité (Thorpe *et al.*, 2001).

Il était aussi important de savoir si ces traitements rapides étaient réservés à des stimuli naturels (animaux, visages) ou s'ils pouvaient être étendus à d'autres catégories. VanRullen & Thorpe ont donc réalisé une expérience utilisant deux catégories d'images (animaux et véhicules). Les tâches proposées aux sujets étaient alternativement de répondre lorsque l'image contenant un animal ou alors lorsqu'elle contenait un véhicule. Chaque lot d'image était donc à la fois une cible dans une partie de l'expérience, puis un distracteur dans l'autre. Les auteurs ont d'abord montré que cette inversion était parfaitement réalisée par les sujets au niveau comportemental, avec des performances équivalentes pour les deux catégories (résultat confirmé par Walther et Fei-Fei, 2007). Ils se sont surtout servis de ce design expérimental pour tester la nature de l'activité différentielle à 150 ms. Plutôt que de soustraire l'activité ERP entre deux catégories qui sont à la fois définies par leurs rôles de cibles et de distracteurs, mais aussi par leurs différences physique intrinsèques, ils ont soustrait l'activité de chaque catégorie lorsqu'elle était cible, avec l'activité provoquée par cette même catégorie lorsqu'elle était distracteur (ce qu'ils ont appelé une différentielle de type 2, VanRullen et Thorpe, 2001). Avec cette opération, la différentielle à 150 ms était conservée, ce qui démontrait sa valeur en tant que réel support de la décision plutôt que manifestation de différence bas-niveau (voir aussi Antal *et al.*, 2000). Malgré quelques objections basées sur la non-corrélation entre cette activité sur un essai donné et le TR observé (Johnson et Olshausen, 2003), cette différentielle semble donc bien être la manifestation d'un processus de catégorisation indépendant des traits physiques de la scène. Une étude IRMf a permis de montrer qu'elle proviendrait des gyri fusiformes et para-hippocampaux du lobe temporal, aires généralement associées à la reconnaissance d'objets (Fize *et al.*, 2000).

Quel rôle joue l'attention dans ces mécanismes ? Pour étudier ces questions, différents protocoles ont été utilisés. Le paradigme de la recherche visuelle, un "Où est Charlie ?" de laboratoire, a permis via un grand nombre d'études de découvrir les attributs de la scène qui pouvaient attirer l'attention automatiquement (on dit alors que ces attributs *pop-out*). Une version simple de ce paradigme consiste par exemple à présen-

Table1 | **Attributes that might guide the deployment of attention**

Undoubted attributes*	Probable attributes‡	Possible attributes§	Doubtful cases‖	Probable non-attributes¶
-Colour[26,27,37,39,40] -Motion[30,56,57] -Orientation[41,42,58–61] -Size (including length and spatial frequency)[27,62,63]	-Luminance onset (flicker)[64,65] -Luminance polarity[21,66] -Vernier offset[67] -Stereoscopic depth and tilt[68–70] -Pictorial depth cues[71–73] -Shape[27,58,74–80] -Line termination[22,81,82] -Closure[26,77,83–85] -Topological status[77,86,87] -Curvature[27,67,88]	-Lighting direction (shading)[51,89] -Glossiness (luster)[52] -Expansion[90,91] -Number[27,81] -Aspect ratio[27]	-Novelty[28,53,92] -Letter identity (over-learned sets, in general)[93–95] -Alphanumeric category[96–99]	-Intersection[5,58] -Optic flow[29,91] -Colour change[64] -Three-dimensional volumes (such as geons)[100,101] -Faces (familiar, upright, angry and so on)[102–108] -Your name[109] -Semantic category (for example, 'animal','scary')[10]

FIGURE 1.10: Les caractéristiques pouvant attirer l'attention de façon automatique. Ceux-ci sont groupés en fonction de leur probabilité à vraiment attirer l'attention. Extrait de Wolfe et Horowitz (2004).

ter à l'écran un nombre variable de stimuli, un seul correspondant à la cible. Cette cible peut être défini selon un attribut simple (rouge parmi les verts) ou une conjonction de ces attributs (rond rouge parmi des ronds verts, des carrés rouges et des carrés verts). Un résultat typique utilisant ce paradigme est que le temps passé à faire une recherche sur un attribut simple va être indépendant du nombre de distracteur. Au contraire, une recherche basée sur une conjonction d'attributs prendra un temps proportionnel au nombre de distracteur. On dit alors que la couleur, dans ce cas, attire l'attention automatiquement, ou plutôt qu'elle *pop-out* (Treisman et Gelade, 1980). Comme on le voit dans la figure 1.10, les attributs capables d'amener un effet *pop-out* semblent être principalement des attributs *simples* : la couleur, le mouvement, l'orientation ou encore la taille (Wolfe et Horowitz, 2004). Comment se situe la détection d'un animal dans ce cadre ? Au premier abord, un animal, comme la plupart des objets réels, consiste en un arrangement complexe de traits et ferait donc parti des stimuli qui nécessitent l'attention. Dans le même temps, les temps de réaction extrêmement rapides supposeraient le contraire. Le traitement de l'animal serait-il donc pré-attentif ? Comme nous l'avons vu, le traitement semble être plutôt parallèle (traitement possible de deux scènes en même temps, Rousselet *et al.*, 2002). Cependant, à partir de quatre scènes, les performances chutent. Le traitement n'est donc pas parallèle au sens strict de la recherche visuelle, où un rond rouge qui pop-out peut attirer l'attention quelque soit le nombre de ronds verts en distracteur. Ainsi, lorsque le nombre de scènes à traiter est multiplié à la façon d'un protocole de recherche visuelle, la pente de temps de réaction en fonction du nombre de distracteurs révèle que la détection d'animal est plus similaire à une recherche de T parmi des L que de T parmi des +. Ce qui signifie que l'animal ne causerait pas d'effet pop-out (VanRullen *et al.*, 2004).

Un autre type de protocole amène à un conclusion différente : la détection d'animal serait pré-attentive. En effet, dans un protocole de double tâche où les sujets doivent réaliser une tâche complexe au centre de l'écran qui nécessite toute leur attention, les sujets restent tout à fait capable de détecter la présence d'un animal en périphérie (Li *et al.*, 2002; Fei-Fei *et al.*, 2005). De même, une étude récente a montré que lorsque l'on

présentait quatre scènes simultanément à un sujet, l'information concernant le contenu de ces scènes pouvait être retrouvée dans l'activité des régions sélectives aux objets (LOC), même si son attention spatiale n'était pas dirigée vers l'image et même s'il ne recherchait pas spécifiquement cet objet (Peelen *et al.*, 2009).

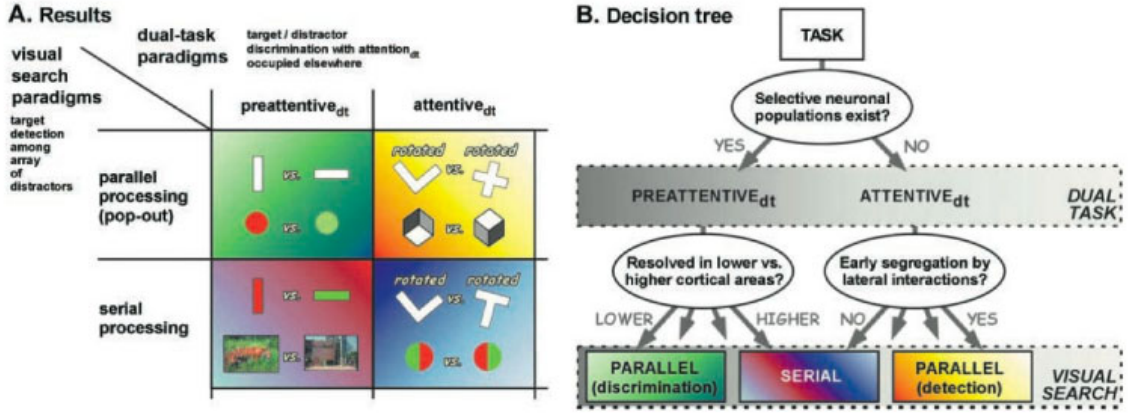

FIGURE 1.11: Processus attentionnels différents pour la recherche visuelle et la double tâche. (A) Deux dimensions indépendantes sont nécessaires pour expliquer ces types les résultats avec les deux protocoles : sériel vs. parallèle pour la recherche visuelle, et pré-attentif et attentif pour la double tâche. Des exemples peuvent être trouvés pour les quatres combinaisons possibles entre ces deux types d'attention. (B) Modèle proposé pour expliquer ces phénomènes. Si une population neuronale sélective existe pour la cible, alors elle peut-être traitée pré-attentivement. Dans le cas d'un traitement pré-attentif, les résultats en recherche visuelle seront ensuite prédits par le niveau de traitement (au sens de la hiérarchie du système visuel). Ces résultats expliqueraient donc pourquoi la détection d'un animal dans une scène naturelle est pré-attentive mais pas pop-out. Extrait de VanRullen *et al.* (2004).

Comment réconcilier ces résultats contradictoires ? VanRullen et al. proposent une dissociation entre mécanismes attentionnels régissant l'aspect sériel/parallèle et ceux concernant les aspects pré-attentifs/attentifs (voir figure 1.11, VanRullen *et al.* (2004)). Cependant, ces deux conclusions (l'animal est traité pré-attentivement et ne pop-out pas) peuvent être relativisées. En effet, en utilisant un protocole de clignement attentionnelle (anglais : attentional blink ; deux cibles sont inclues dans un flux d'image RSVP, la deuxième n'est pas perçue lorsqu'elle est trop proche de la première), il a été montré que la détection d'animal souffrait aussi de ce phénomène. Un certain type d'attention pourrait donc aussi être requis (Evans et Treisman, 2005). Les résultats en recherche visuelle peuvent aussi être interprétés différemment. Les scènes étaient ici considérées comme les stimuli et non pas les objets. Il se peut alors que la difficulté de la tâche vienne de la multitude de scènes à traiter en parallèle, qui écraserait l'effet *pop-out* de l'animal au sein d'une même scène. Une autre approche serait de considérer une scène visuelle comme un affichage de recherche visuelle à part entière, chaque région de la scène jouant le rôle de distracteur, et l'objet d'intérêt étant la cible. Dans ce cadre, il se pourrait que l'animal pop-out. Une étude a indirectement accrédité cette

hypothèse. Elle comparait la détection d'animal (vs. objet) lorsqu'ils étaient présentés en contexte (dans une scène naturelle) ou sur un fond gris uniforme (Joubert *et al.*, 2008). On peut donc considérer cette deuxième condition comme "sans distracteur" au sens de la recherche visuelle, alors que la première serait une condition avec distracteurs (les autres régions de la scène). La différence de temps de réaction entre les deux conditions n'était pas significative. Il semblerait donc bien que, dans une certaine mesure, l'animal pop-out de la scène. L'idée même de pop-out, définie sur la base d'une non augmentation du temps de réaction en fonction du nombre de distracteurs, reste donc tout de même difficile à adapter aux scènes naturelles, car le nombre de distracteurs ne peut pas être défini précisément au sein d'une même scène.

1.4.2 Encore plus vite : la tâche de choix saccadique

Nous avons donc vu que la détection d'objet dans une scène complexe pouvait être beaucoup plus rapide que ce que la plupart des modèles peuvent prédire, avec des temps de réaction très courts qui semblent être supportés par une activité différentielle dans les aires temporales 150 ms après l'affichage des stimuli. En 2006, une étude a été encore plus loin dans la recherche des toutes premières réponses sélectives possibles. Profitant de la possibilité de traiter deux images en parallèle (Rousselet *et al.*, 2002), Holle Kirchner et Simon Thorpe ont mis au point un nouveau protocole se basant sur une nouvelle modalité de réponse : les saccades oculaires (Kirchner et Thorpe, 2006). C'est ce protocole qui aura servi de base à l'ensemble des résultats collectés au cours de cette thèse. Il se déroulait de la façon suivante (Figure 1.12) :

- Affichage d'une croix de fixation pendant 800 à 1600 ms (pseudo-aléatoire à chaque essai).
- Intervalle de 200 ms (communément appelé gap et connu pour accélérer l'initiation des saccades oculaires (Fischer et Weber, 1993).
- Affichage de 2 images (une cible et un distracteur) sur la droite et la gauche de l'écran pendant 20 ms (Kirchner et Thorpe, 2006), puis 400 ms par la suite [1].
- Pause inter-essai de 1 seconde.

Dans cette expérience de discrimination animal/non-animal, et grâce à ce protocole, les sujets produisaient des réponses encore bien plus rapides que ce qui avait été enregistré auparavant. Et de façon extrêmement surprenante, les premières réponses sélectives (le temps de réaction minimum) apparaissaient même avant l'activité différentielle à 150 ms, censée être le support à la décision pour les réponses de catégorisation rapide. Les sujets les plus rapides avaient ainsi des réponses sélectives dès 120/130 ms. Si l'on

1. En effet, l'affichage de 20 ms, hérité du protocole go/no-go, était initialement prévu pour éviter que les sujets aient le temps d'explorer la scène. Avec le protocole de choix saccadique, nous enregistrons justement les mouvements oculaires, cet affichage limité était donc obsolète. De plus, comme l'ont montré des études préliminaires, ainsi que le premier article de cette thèse (voir section 2.2), un affichage plus long tend à raccourcir les temps de réaction.

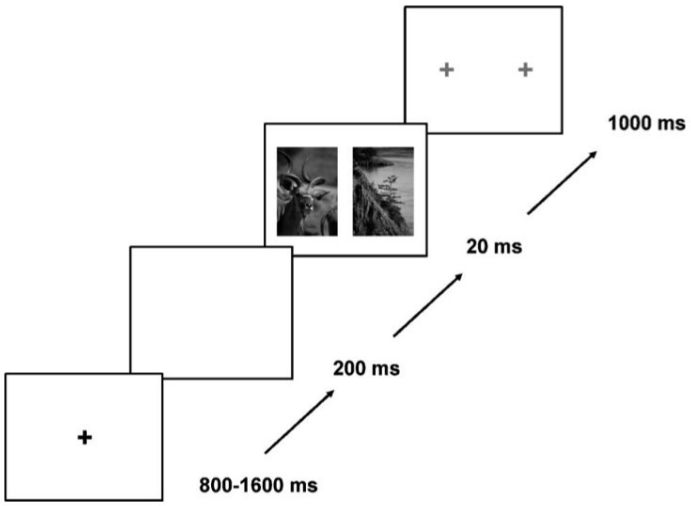

FIGURE 1.12: **Protocole de choix saccadique**. Déroulement d'un essai typique. **(1)** Affichage d'une croix de fixation pendant 800 à 1600 ms (pseudo-aléatoire à chaque essai). **(2)** Intervalle de 200 ms (communément appelé gap et connu pour accélérer l'initiation des saccades oculaires (Fischer et Weber, 1993). **(3)** Affichage de 2 images (une cible et un distracteur) sur la droite et la gauche de l'écran pendant 20 ms. **(4)** Pause de 1 sec en attendant de recommencer le prochain essai. Extrait de Kirchner et Thorpe (2006).

considère qu'environ 20/30 ms sont nécessaires depuis la prise de décision jusqu'à l'initiation effective d'une saccade (Stanford *et al.*, 2010), cela signifie que l'information était disponible seulement 100 ms environ après l'affichage des images. De nombreuses analyses sur les statistiques des images ont été faites pour écarter certaines explication bas-niveau comme des différences de luminance, de contraste, d'énergie, etc. Pour ceci, les auteurs calculaient un grand nombre de statistiques pour chaque image, enlevaient ensuite des données les images ayant des valeurs extrêmes sur ces caractéristiques, et enfin refaisaient les mêmes analyses pour voir si un changement était notable. Ce ne fut pas le cas. Comment expliquer alors que l'information puisse être exploitée si vite par le système oculomoteur ? Deux hypothèses sont possibles. Une première possibilité est que l'information prenne un raccourci (peut-être même via des structures sous-corticales) pour atteindre directement les zones visuo-motrices (FEF ou LIP, voir section 1.6 pour plus de détail sur ces structures). Du fait de la complexité des scènes naturelles, les auteurs ont cependant privilégié une autre alternative, consistant en une prise d'informations avant d'atteindre les aires habituellement considérées comme le siège de la reconnaissance d'objets. Ainsi, il se pourrait que les sujets détectent des traits simples caractéristiques des animaux (par exemple ceux de V4), plutôt que l'information de neurones de haut-niveau comme ceux d'IT. Ce modèle est détaillé dans la Figure 1.13. Un support physiologique possible pour ces réponses ultra-rapides a été proposé par la suite. Avec des enregistrements EEG intra-craniaux chez des patients épileptiques, les auteurs ont montré que le signal dans FEF pouvait devenir sélectif à la catégorie d'image (non pas la catégorie entre images naturelles, mais entre une image en damier,

une image colorée, une image naturelle) seulement 45-60 ms après la présentation des stimuli (Kirchner *et al.*, 2009). Cette activation rapide pourrait se faire selon différents chemins (sous-corticaux ou corticaux) mais semble être un support plausible pour les résultats observés dans une tâche de choix saccadique.

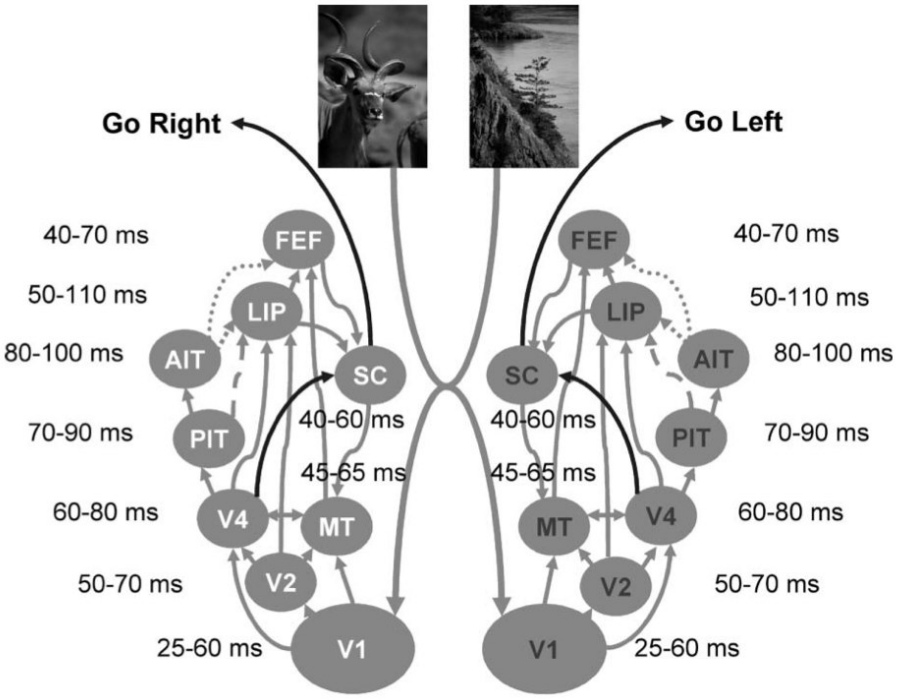

FIGURE 1.13: Modèle hypothétique des activations cérébrales de la voie ventrale lors de la tâche de choix saccadique. Les estimations de latence données ici sont basées sur des données électrophysiologiques chez le singe, il faut donc probablement ajouter quelques millisecondes pour faire l'équivalence avec les humains. Le premier nombre pour chaque étape correspond aux latences des premières réponses observées, le second correspond à une valeur moyenne. Les flèches provenant de AIT et PIT sont en pointillés car l'existence de ces connections est hypothétique. Extrait de (Kirchner et Thorpe, 2006).

Plusieurs autres études ont utilisé ce protocole chez l'humain, mais aussi chez le primate (ceux-ci étant encore plus rapide, Girard *et al.*, 2008). Il a aussi été démontré que les performances des sujets humains pour la discrimination animal-non-animal était particulièrement résistante à la rotation (Guyonneau *et al.*, 2006), ainsi qu'à diverses dégradations de la luminance des images (floutage, inversion, discrétisation, Nandakumar et Malik, 2009). Une autre étude a consisté, toujours à partir d'une tâche animal/non-animal à comparer les différents protocoles (réponse manuelle : go/no-go, yes/no et choix entre les deux côtés ; réponse saccadique : tâche de choix saccadique Bacon-Macé *et al.*, 2007). En utilisant un paradigme de masquage, les auteurs ont ainsi montré que la phase de traitement sensoriel précoce (SOA inférieur à 40 ms) semblait être similaire entre les différents protocoles.

Ce nouveau protocole ouvrait donc des perspectives très intéressantes, en offrant un accès à une phase très précoce des traitements visuels chez l'humain. Il a donc été

le point de départ de mon travail de thèse, qui a consisté à étudier les mécanismes qu'il impliquait, tout en l'utilisant pour tester de nouvelles hypothèses sur ces traitements ultra-rapides. Cependant, avant de passer à la partie expérimentale de cette thèse, qui s'inscrit dans la continuité des travaux présentés dans cette section, nous allons d'abord nous intéresser à deux domaines qui deviennent fondamentaux lorsque l'on utilise ce protocole de choix forcé saccadique : le guidage des mouvements oculaires et la prise de décision saccadique.

1.5 Qu'est ce qui attire le regard ?

L'objectif de cette thèse n'est pas explicitement de comprendre le type d'information qui attire le regard, mais de se servir des mouvements oculaires comme d'un outil donnant accès aux toutes premières étapes du traitement visuel. Cependant, de nombreuses études ont étudié de façon précise la façon dont nous bougeons les yeux lorsque nous explorons des scènes visuelles, et de nombreux modèles ont été proposés pour expliquer ce qui attirait le regard. Ces études, malgré un objectif de départ légèrement différent, peuvent nous fournir un grand nombre d'informations. Les principaux modèles d'exploration de scènes seront donc détaillés, avant d'étudier plus précisément ce qu'ils impliquent sur la nature des informations guidant la toute première saccade.

Essayer de comprendre ce qui peut attirer le regard dans une scène visuelle intéresse depuis très longtemps les scientifiques (ainsi que les publicitaires). Ceci a commencé à être étudié dès le 19ème siècle par observation directe, ce qui ne permettait pas de faire de vraie recherche scientifique. La première étude à enregistrer et à analyser les mouvements oculaires de sujets explorant des images date de 1935 (Buswell, 1935). Buswell avait mis au point un système ingénieux utilisant des faisceaux de lumière qui venaient se refléter sur la cornée, cette réflexion pouvant ensuite être enregistrée sur un film. Ceci permettait pour la première fois d'enregistrer la position du regard de façon non-invasive. Il testa ainsi 200 sujets regardant 55 images différentes allant de peintures de grands peintres à des images de tapisseries en passant par des photographies de design d'intérieur. En 1967, Yarbus publia un livre qui deviendra une des références les plus citées dans le domaine de l'exploration oculaire (Yarbus, 1967). Il y aborde déjà un grand nombre de questions qui sont encore d'actualité, et y pose les bases de l'étude des mouvements oculaires. Son observation la plus connue a été de démontrer l'influence de la tâche sur la façon dont les sujets peuvent explorer une même photographie (voir Figure 1.14).

Une grande partie des études concernant la génération des mouvements oculaires s'est basée sur l'utilisation de stimuli bas-niveau. Le modèle majeur, rendant compte de la plupart des résultats expérimentaux observés dans ce cadre, est celui de Findlay

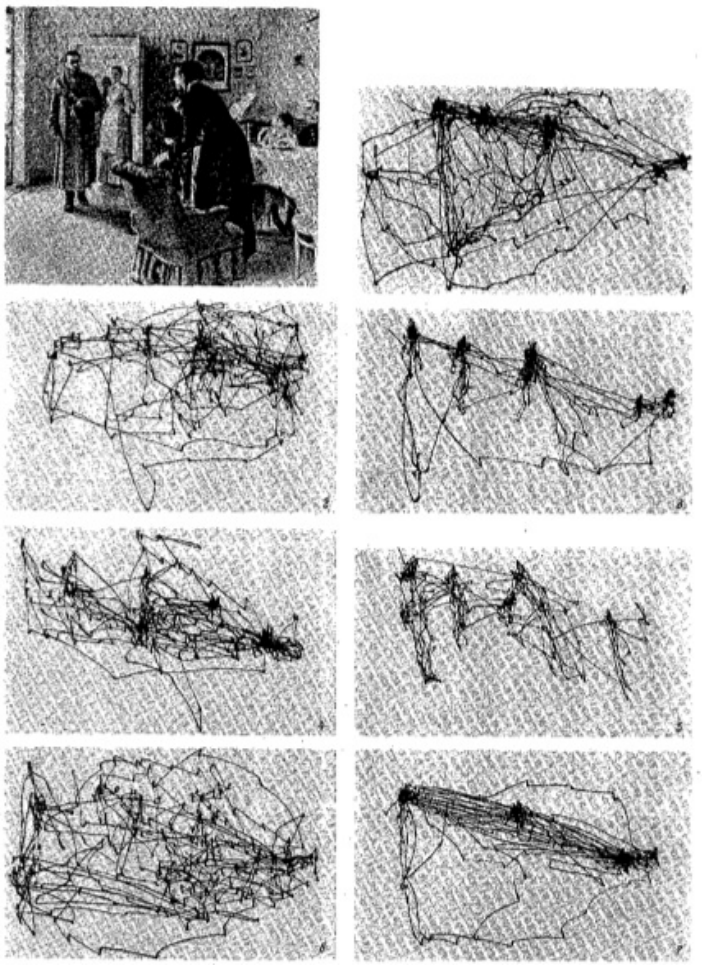

FIGURE 1.14: 7 enregistrements des mouvements oculaires produits par le même sujet devant la même image mais avec des tâche différentes (Yarbus, 1967). 1) Exploration libre 2) Estimer la condition financière de la famille 3) Donner l'âge des personnages 4) Imaginer ce que la famille faisait avant que le visiteur entre dans la pièce 5) Mémoriser les vêtements portés par les personnages 6) Mémoriser la position des objets et des personnes dans la pièce 7) Estimer depuis combien de temps ce visiteur n'était pas venu leur rendre visite.

et Walker (1999). Il permet d'expliquer la plupart des effets observés dans les protocoles classiques de mouvement oculaires : accélération des saccades lorsqu'un gap est placé entre la fin de la fixation et l'apparition des stimuli, bi-modalité fréquente des distributions de réponses (saccades *express* vs. normales), ralentissement des saccades lorsque plusieurs stimuli sont présentés simultanément, et enfin ralentissement des réponses dans le cas d'une tâche d'anti-saccade. Ce modèle rend compte à la fois des aspects temporels (quand déclencher la saccade) et spatiaux (vers où). Cependant, il ne rend pas compte de l'ensemble des phénomènes observés dans la littérature, notamment dans les conditions que sont la lecture (Brysbaert *et al.*, 2005) ou l'exploration de scènes naturelles que je vais maintenant détailler plus en profondeur.

Dans le cadre du traitement des scènes naturelles, les travaux récents étudient plus particulièrement les différents attributs d'une image qui vont nous amener à bouger les

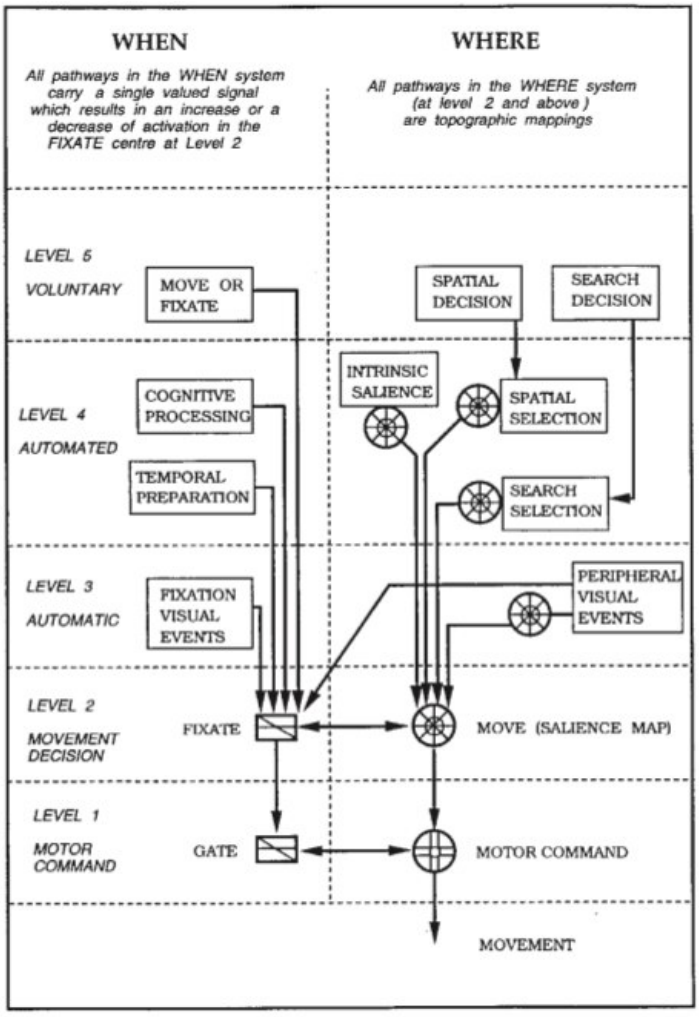

FIGURE 1.15: **Modèle de génération de saccades de Findlay & Walker (1999).** Selon eux, la localisation et le moment de déclenchement de la saccade sont deux processus relativement indépendants, ce qui est symbolisé par les deux voies (à droite et à gauche) du modèle. Trois paramètres viennent influencer la génération de saccades : la saillance bas-niveau, la sélection spatiale et la tâche à réaliser. Ces deux derniers paramètres pouvant avoir des composantes volontaires. Une hypothèse forte de ce modèle, et qui concernera tout particulièrement les résultats de cette thèse, est que la reconnaissance d'objet prendrait trop de temps pour pouvoir jouer un rôle significatif dans le guidage des mouvements oculaires.

yeux. Dans la lignée du modèle de Findlay et Walker 1999, deux aspects principaux peuvent influencer le guidage des mouvements oculaires : le regard peut être *attiré* vers un endroit précis de la scène par les propriétés physiques de cette région, ou alors *envoyé* vers un endroit par des facteurs cognitifs plus haut-niveau relatif à nos connaissances préalables ou nos intentions (Henderson, 2007). Ces deux facteurs peuvent être compris assez intuitivement : notre regard va être *attiré* par une forte lumière, mais *envoyé* vers la table si l'on cherche une bouteille.

Puisque les facteurs physiques des stimuli sont plus faciles à aborder que les facteurs cognitifs descendants, une grande partie des recherches s'est attachée à l'étude des premiers cités. Une façon d'aborder ce problème a notamment été de comparer

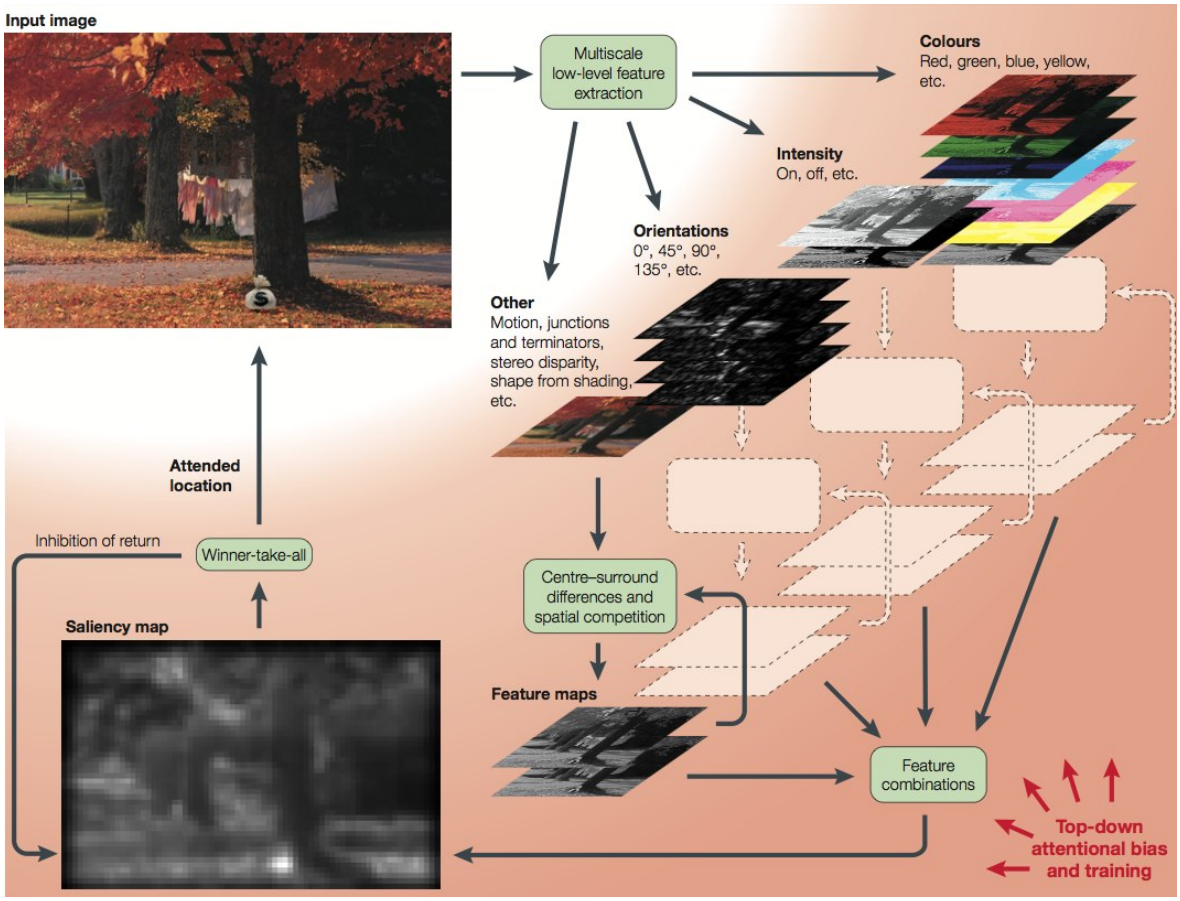

FIGURE 1.16: Le modèle de saillance bas-niveau d'Itti & Koch, extrait de Itti et Koch (2001). L'image est traitée dans plusieurs canaux indépendants qui vont chacun s'occuper d'une caractéristique différente (orientation, couleur, luminance, mouvement par exemple). Une carte est ensuite crée pour chaque canal, puis elles sont combinées pour donner lieu à la carte de saillance. La zone la plus activée de cette carte correspondra à la première fixation, et ainsi de suite.

les caractéristiques des régions fixées à celles des régions non-fixées. Nos yeux se dirigeraient ainsi préférentiellement vers les régions contenant des bords, ou celles étant plus contrastées. Une autre méthode, basée sur la théorie d'intégration de traits (Treisman et Gelade, 1980), tente de prédire les fixations des sujets (Itti *et al.*, 1998). Ici, l'image est traitée par différents modules, chacun responsable d'une propriété (couleur, contraste, luminance, mouvement par exemple). Les données de chacun de ces traitements sont ensuite combinées pour former une *carte de saillance*, où les zones les plus susceptibles d'attirer le regard sont les plus actives (voir Figure 1.16 pour un exemple).

Un des problèmes majeurs avec ce type d'approche est qu'elle n'utilise que des corrélations entre statistiques d'images et fixations, ou régions sélectionnées et fixations. L'effet de causalité entre les deux n'est pas avéré. Ainsi, la présence d'un objet dans le scène va forcément ajouter des bords, ou amener une variation de contraste différente. La question est donc de savoir si ce qui va attirer le regard est cet aspect bas-niveau, ou bien plutôt l'objet en lui-même. Des sujets à qui l'on demande d'annoter

les images utilisées dans ce type d'expérience vont considérer comme plus informative les régions statistiquement différentes (Henderson *et al.*, 2007). De nombreux autres résultats viennent mettre en doute la réelle pertinence de la saillance bas-niveau. Ainsi, son effet peut être inversé (exemple : les contrastes faibles deviennent plus saillants) lorsque cela devient pertinent pour la tâche (Einhäuser *et al.*, 2008a). De même, elle s'avère beaucoup moins prédictive que la simple position des objets (Einhäuser *et al.*, 2008b), et ses performances de prédiction chutent largement si la scène contient des visages (Birmingham *et al.*, 2009). Le couplage avec un algorithme de détection de visage permet ainsi d'augmenter fortement la qualité de la prédiction (Cerf *et al.*, 2009).

Il semble donc que de nombreux facteurs autres que la pure saillance bas-niveau puissent influencer les mouvements oculaires. En effet, en dehors des caractéristiques visuelles physiques de la scène, ceux-ci peuvent aussi être influencés par des facteurs internes qui vont *envoyer* les yeux dans des régions spécifiques. Les deux facteurs de ce type les plus étudiés sont la recherche spécifique d'une cible et l'influence du contexte. Lorsque l'on recherche une cible spécifique, on peut imaginer qu'une représentation préalablement activée en mémoire de travail va servir de modèle que l'on cherchera dans la scène présentée. Les régions de la scène les plus proches du modèle seront alors sélectionnées pour la fixation (Malcolm et Henderson, 2009; Rao *et al.*, 2002; Zelinsky, 2008). Ce type de processus peut être facilement combiné à l'influence de la saillance bas-niveau : la recherche de cible peut se traduire par une modulation des traitements ascendants (Maunsell et Treue, 2006) afin de biaiser la carte de saillance comme dans le modèle de Rao et le modèle SUN (Saliency Using Natural statistics) (Rao *et al.*, 2002; Kanan *et al.*, 2009). C'est ce qu'il se passe par exemple lorsque l'on cherche un ami sur les pistes parmi les autres skieurs. Si je sais qu'il porte une veste verte et un pantalon bleu, je peux alors filtrer les informations visuelles ascendantes pour que seulement les personnes affichant ces caractéristiques soient sélectionnées. Cependant, toutes les connaissances ne peuvent pas venir moduler les traitements ascendants aussi facilement. C'est la cas notamment des informations contextuelles spatiales.

Quand le système visuel utilise le contexte de la scène, ses connaissances préalables sont utilisées pour identifier les régions de la scène les plus susceptibles de contenir l'objet cible (Torralba *et al.*, 2006). L'idée ici n'est pas de venir directement moduler la carte de saillance, mais d'effectuer un traitement parallèle qui servira ensuite à modifier le résultat du traitement ascendant. Le modèle de guidage contextuel de Torralba en est la réalisation la plus aboutie (voir Figure 1.17). Cette étude montre que les informations contextuelles s'avèrent plus pertinentes que la saillance bas-niveau pour prédire les mouvements oculaires. Cependant, c'est bien la combinaison des deux qui permet de se rapprocher le plus des patrons d'exploration humains. Les modèles utilisant donc à la fois les informations ascendantes (saillance) et descendantes (contexte ou modèle de cible) ont des performances de prédiction qui dépassent largement ce que font les

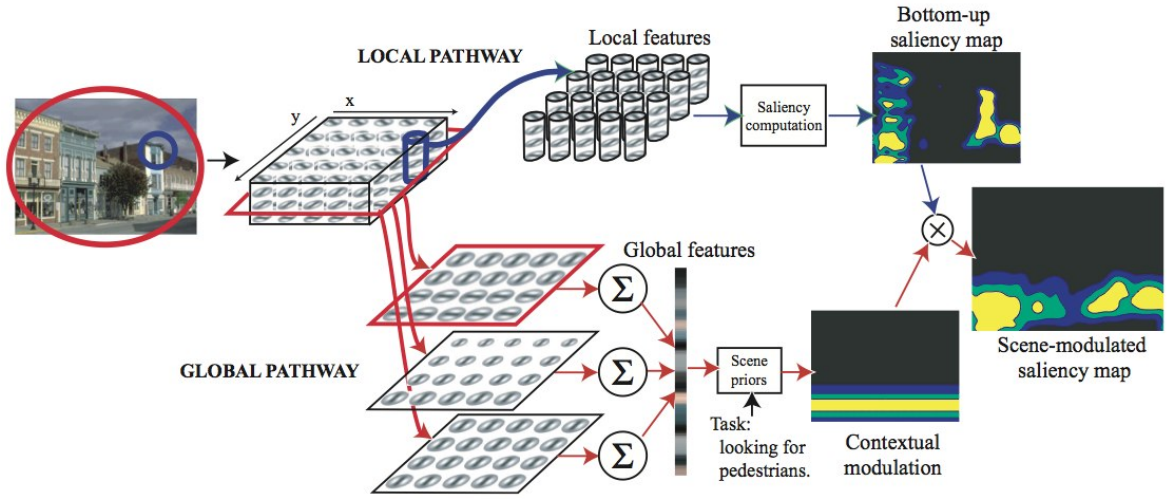

FIGURE 1.17: Le modèle de guidage contextuel de Torralba *et al.* (2006) qui intègre la saillance de l'image et les *a priori* sur la scène. L'image est analysée par deux voies parallèles qui partagent une première étape identique dans laquelle l'image traverse un ensemble de filtres orientés multi-échelle. La voie locale (en haut), traite chaque région indépendamment et sert à créer une pure carte de saillance bas-niveau. Le voie globale (en bas), extrait les statistiques globales de l'image qui peuvent être utilisées pour la reconnaissance de scène, et va venir contraindre la voie locale vers les zones les plus susceptibles de contenir les objets d'intérêt.

modèles simples (Torralba *et al.*, 2006; Kanan *et al.*, 2009; Rao *et al.*, 2002). La toute dernière version du modèle de Torralba, intégrant à la fois la saillance bas-niveau, les influences contextuelles, et les modèles de cible, s'avère être pour l'instant le modèle le plus abouti de guidage oculaire dans les scènes naturelles (Ehinger *et al.*, 2009).

Dans cette thèse, je me suis particulièrement intéressés aux traitements visuels les plus précoces. Lors de l'exploration d'une scène visuelle pendant plusieurs secondes, de nombreux processus ont le temps de rentrer en jeu. Il se pourrait donc qu'ils ne soient pas tous mis en jeu au même moment. Ce phénomène se manifeste notamment par un taux de cohérence entre les sujets beaucoup plus important pour la première saccade, et qui décroît ensuite largement (Parkhurst *et al.*, 2002; Carmi et Itti, 2006). Les différents auteurs semblent être à peu près tous d'accord sur un rôle plus important de la saillance bas-niveau pour la détermination du tout premier point de fixation, les autres processus venant jouer un rôle par la suite (Itti, 2005; Parkhurst *et al.*, 2002; Henderson et Hollingworth, 1999; Tatler *et al.*, 2005). Cependant, cet effet n'a pas été répliqué dans deux études récentes, qui ne montrent pas d'effet différentiel sur la première saccade (Einhäuser *et al.*, 2006; Kayser *et al.*, 2006).

1.6 Sélection et décision

Nous venons de voir ce qui pouvait guider les mouvements oculaires lorsque les sujets exploraient des scènes. Cependant, dans la plupart des expériences de psycho-

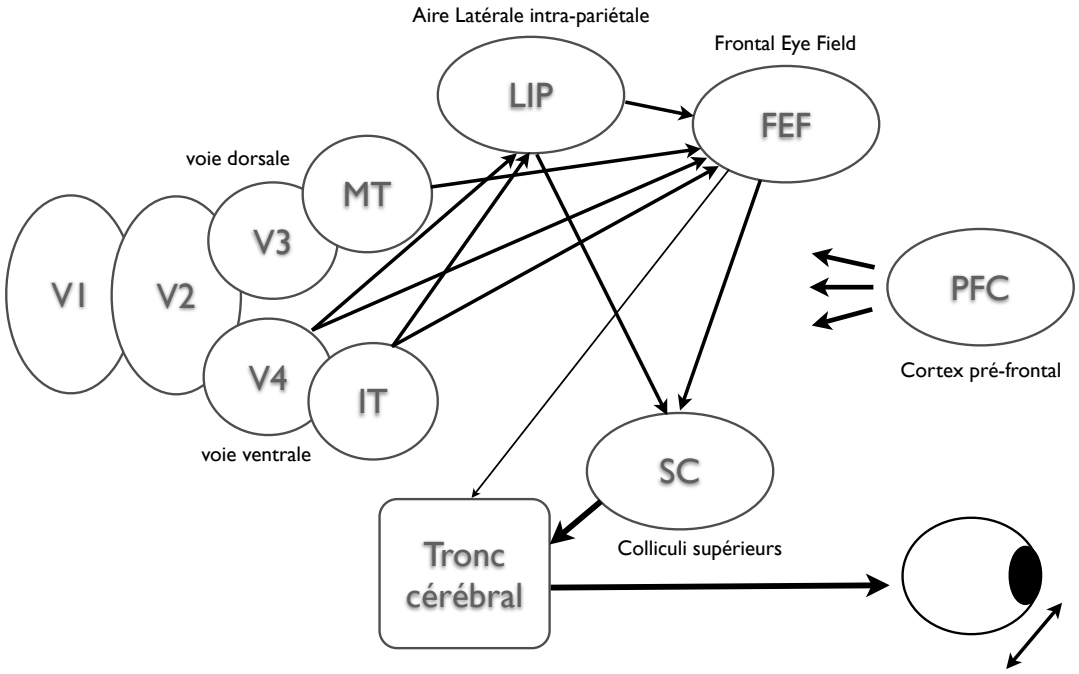

FIGURE 1.18: Schéma des différentes régions impliquées dans le contrôle des saccades et de leurs connexions (entre-elles, avec les aires purement visuelles, et avec le tronc cérébrale où se trouve les noyaux des motoneurones oculaires).

physique, les sujets doivent effectuer une tâche, c'est à dire rechercher une cible qui se trouve en compétition avec un ensemble de distracteurs. Les processus décisionnels permettant cette sélection vont donc jouer un rôle majeur. De nombreux articles et revues sont parus en neurosciences et psychologie dans le domaine de la prise de décision. Je tente ici d'en faire un résumé basé essentiellement sur les études en électrophysiologie chez le singe initiées par les équipes de Shadlen et Newsome dans les années 90.

La plupart des études sur la décision perceptive en neurosciences ont été menées via les réponses oculaires saccadiques. Ceci est très pratique pour nous car c'est justement cette modalité de réponse qui va être utilisée par la suite dans nos travaux expérimentaux. Dans le section précédente, nous avons vu que la saillance des différents stimuli dans notre champ visuel pourrait être représentée sous la forme d'une carte topographique (voir section 1.5).

Trois aires cérébrales sont des candidates probables pour être le siège de ces cartes :

- L'aire oculomotrice frontale (anglais : Frontal Eye Field, FEF) (Hanes et Schall, 1996)
- Le cortex intra-pariétal latéral (LIP) dans le cortex pariétal postérieur (Roitman et Shadlen, 2002)
- Les colliculi supérieurs (SC) (Horwitz *et al.*, 2004b,a)

Il se trouve que chacune de ces aires est aussi impliquée plus ou moins directement

dans la motricité des mouvements oculaires (voir Figure 1.18). C'est pourquoi on les appelle généralement des aires visuo-motrices. Pour résumer, lorsque plusieurs stimuli sont présentés en même temps dans le champ visuel, ils sont alors représentés dans des cartes topographiques. Ceci peut-être vu comme une carte de chaleur, avec des points chauds correspondant à chacun des stimuli. L'activité des neurones activés par la cible va ensuite diverger de celle des distracteurs, et devenir la cible de la saccade (Glimcher, 2001; Schall, 2001). Autrement dit, la carte de chaleur s'est modifiée, et la position qui est restée active est celle vers laquelle le sujet a finalement exécuté sa saccade. On passe donc ici d'une simple représentation de l'ensemble des stimuli à une sélection de celui qui devient la cible de la saccade. Comment ce processus de sélection s'est-il effectué ?

Pour répondre à cette question, l'équipe de Schall a élaboré une tâche dans laquelle le singe regardait un écran avec 8 stimuli, dont 7 identiques. Celui-ci n'était récompensé que s'il effectuait une saccade vers le stimulus différent des autres. Lorsque les 8 stimuli s'affichent, ils sont tous traités par le système sensoriel pour être représentés ensemble dans la carte topographique FEF qui est étudiée ici. En enregistrant directement l'activité des neurones de FEF chez le singe, les auteurs ont montré que les taux de décharge des neurones associés aux différentes alternatives étaient similaires pendant les 80 premières millisecondes. Au delà, le signal correspondant à la cible surpassait largement les signaux associés aux 7 autres distracteurs. Ils ont aussi montré que la pente de cette augmentation d'activité était corrélée à la vitesse du temps de réaction. Autrement dit, plus l'activité correspondant à la cible augmentait vite, plus la réponse comportementale du singe était rapide. En revanche, l'intervalle de temps constant entre le dépassement du seuil et le mouvement oculaire confirme l'existence d'un seuil (voir Figure 1.19).

Ces résultats montrent que l'activité des neurones dans FEF n'est pas seulement de coder les stimuli visuels. Le patron d'activité globale va aussi décider vers où vont se diriger les yeux. Mais l'on a vu aussi que la façon dont l'activité des neurones évoluait ressemblait à une intégration d'informations. En effet, il semble que les réponse saccadiques rapides correspondaient spécifiquement à une augmentation rapide de l'activité des neurones de cette aire. Il semble donc qu'en plus d'être une aire visuo-motrice, FEF est aussi une aire décisionnelle.

Pour mieux comprendre les aspects décisionnels dans ces aires visuo-motrices, les équipes de Newsome et Shadlen ont essayé de comprendre le lien entre l'information perceptive contenue dans les neurones sensoriels et l'activité des neurones dans ces aires visuo-motrices. Dans ces expériences, les singes regardaient un nuage de point en mouvement. À chaque essai, un sous-ensemble de ces points bougeait de manière cohérente dans une direction précise. Le singe recevait une récompense lorsqu'il signalait correctement dans quelle direction bougeait ce sous-ensemble de points, en effectuant une saccade dans cette même direction. On sait que les neurones sensoriels impliqués

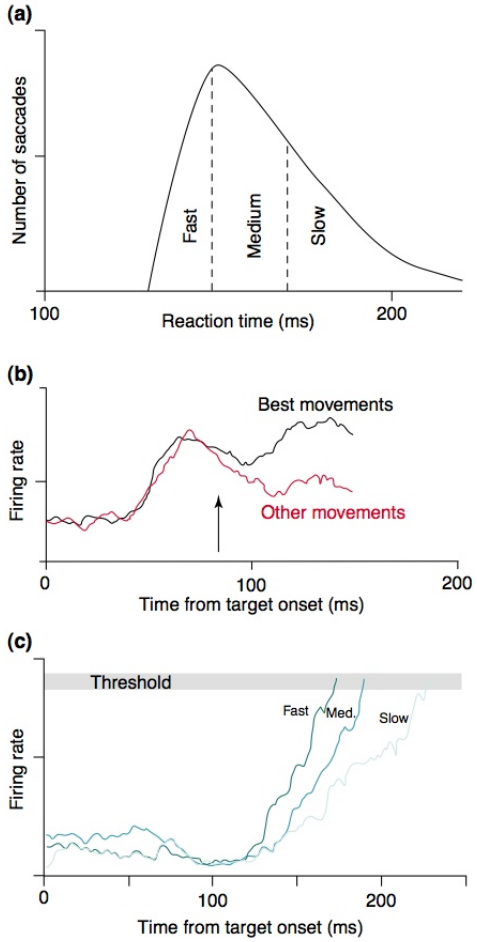

FIGURE 1.19: Les résultat expérimentaux dans FEF par l'équipe de Schall, figure extraite d'une revue de Glimcher (2001). (a) Les saccades ont été séparées en différents groupes selon leur vitesse (rapide, intermédiaires et lentes). (b) L'activité correspondant au choix de la cible diverge de l'activité associée aux autres distracteurs après 80 ms. (c) Les temps de réaction de la saccade sont directement dépendants de la pente de l'augmentation d'activité.

dans ce type de tâche se trouve dans l'aire MT, dont les neurones codent différentes orientations du mouvement. En décodant l'activité dans cette aire, on peut ainsi savoir la direction du mouvement présenté au singe (Kamitani et Tong, 2006).

En 1996, l'équipe de Shadlen a proposé un modèle de décision visuelle saccadique pour cette tâche. Dans ce modèle, le mouvement des points est analysé par les neurones de MT. La sortie globale de MT contient donc à chaque instant donné, une estimation du mouvement des points pour chaque position dans le champ visuel. En moyennant l'activité des neurones de MT sur deux secondes et pour chaque direction et chaque position, ils étaient alors capable de prédire dans quelle direction la future saccade allait se produire (voir Figure 1.20) (Shadlen *et al.*, 1996).

La question est alors de savoir par quelle procédé certaines aires du cerveau vont décoder cette activité sensorielle pour décider comment le singe va répondre. L'enregistrement des neurones du cortex pariétal postérieur (LIP) a montré que leurs patrons d'activité correspondaient parfaitement au rôle d'intégrateur (Shadlen et Newsome,

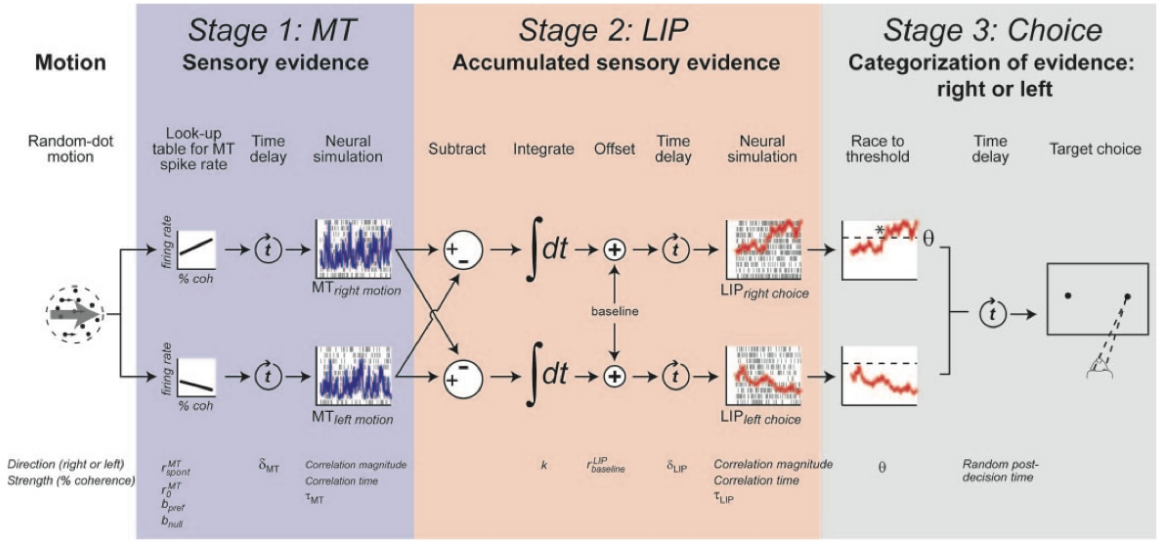

FIGURE 1.20: Le modèle de décision perceptive de Mazurek *et al.* (2003). L'activité des neurones de MT spécifique à chaque direction est moyennée. Cette information est intégrée temporellement pour estimer la direction majoritaire des points. Information qui permet ensuite d'initier la saccade dans la bonne direction.

1996; Roitman et Shadlen, 2002). À partir de la présentation des stimuli et jusqu'à la réponse, ces neurones augmentent ainsi graduellement leur taux de décharge. Ce qui suggère qu'ils sont en train d'intégrer au cours du temps les informations des neurones sensoriels auxquels ils sont connectés (Gold et Shadlen, 2007; Heekeren *et al.*, 2008; Schall, 2001). Une réponse comportementale est alors initiée lorsque le taux de décharge du neurone intégrateur correspondant dépasse un certain seuil (Roitman et Shadlen, 2002). Les différents taux de décharge permettaient donc de prédire où le singe allait bouger les yeux. De plus, si la proportion de points bougeant de manière cohérente était augmentée, les taux de décharge de ces mêmes neurones augmentaient eux aussi plus rapidement, ce qui est exactement le comportement que l'on attendrait de neurones jouant le rôle d'intégrateur des informations sensorielles. Dans ces expériences utilisant des stimuli en mouvement et des réponses oculaires, l'aire LIP constitue donc une interface entre les systèmes sensoriels (MT d'où elle reçoit des projections) et les systèmes moteurs (FEF vers laquelle elle projette). La relation de causalité entre l'activité des neurones de LIP et la réponse comportementale a été attestée par une étude de microstimulation (Hanks *et al.*, 2006).

Il semble donc que les régions impliquées dans la représentation de variables de décision seraient aussi celles effectuant la décision et planifiant l'action, ces régions sensori-motrices *liraient* l'activité des populations de neurones sensoriels. Elles coderaient ensuite sélectivement les différents choix possibles de telle façon que le niveau d'activation de leurs neurones indiquerait la réponse du sujet (Gold et Shadlen, 2007). Dans ce cadre, il semble que le rôle du cortex pré-frontal serait de moduler le seuil de

déclenchement de la réponse en fonction de la difficulté de la tâche (cortex pré-frontal postérieur médian, pmPFC) (Bogacz *et al.*, 2009; Heekeren *et al.*, 2008).

Pour résumer, il semble donc que les trois structures que sont FEF, LIP et SC participent toutes à la représentation et à la sélection de cible, ainsi que la génération d'un mouvement oculaire vers cette cible. Ces trois structures sont largement interconnectées. Elle ne sont cependant pas uniformes et comprennent ainsi différents types de neurones. Ceux-ci peuvent être classés en deux populations principales : les neurones à réponse tonique aux stimuli visuels et sans modulation saccadique (les neurones *visuels*), et ceux très peu modulés par les stimuli visuels mais pour lesquels on note une forte augmentation de décharge avant la production d'une saccade (les neurones *de mouvement*). Les neurones *visuels* se trouvent dans les trois structures, alors que les neurones de mouvement sont très peu présents dans LIP. Chez le singe en train de faire une recherche visuelle, les neurones toniques modulent leur activité en fonction de la tâche (ceci a été constaté dans les trois structures). Ce processus de sélection est indépendant de la production d'une saccade. Les neurones toniques sont donc supposés représenter la pertinence d'un objet dans son champ récepteur et seraient la source d'information pour les neurones de *mouvement* (Purcell *et al.*, 2010). Les neurones de mouvement dans FEF et SC déclenchent une saccade lorsque leur taux de décharge atteint un seuil (Hanes et Schall, 1996; Ratcliff *et al.*, 2007). Le temps non-décisionnel pour que l'activité commence à être accumulé, ainsi que la vitesse d'accumulation permet d'expliquer la variabilité des temps de réaction (Hanes et Schall, 1996; DiCarlo et Maunsell, 2005). Une façon de modéliser le processus de prise de décision perceptive pourrait donc être de considérer les neurones visuels (toniques) de FEF comme les indices perceptifs et les neurones de mouvement de FEF comme les accumulateurs de ces indices (Purcell *et al.*, 2010). Le chapitre 4 fera le lien entre ces recherches neurophysiologiques et les modèles mathématiques de prise de décision.

Nous avons donc vu que les neurones de ces aires visuo-motrices encodent la cible correspondant à la tâche. Quels sont alors les mécanismes permettant de changer cette tâche ? Par exemple dans les expériences de ce mémoire (voir chapitre 2), les sujets devaient réaliser des saccades soit vers les visages, soit vers les véhicules. Sachant que les neurones de FEF et LIP encodent la cible de la tâche, un mécanisme plastique qui soit capable de modifier la façon dont ils vont lire l'information dans les populations sensorielles est donc nécessaire. Une étude récente suggère que cette modulation se ferait directement au niveau des populations sensorielles (Ferrera *et al.*, 2009).

1.7 Problématique

Les expériences de classification d'images naturelles utilisant un protocole de choix saccadique ont montré que les sujets humains pouvaient initier des saccades vers l'image contenant un animal dès 120-130 ms (Kirchner et Thorpe, 2006). Ces réponses très précoces apparaissaient donc avant même l'activité différentielle à 150 ms, qui est généralement considérée comme le support neuronal de la catégorisation rapide. De plus, ces réponses comportementales précoces apparaissent sensiblement au même moment que les premières réponses neuronales sélectives dans les zones cérébrales associées à la reconnaissance d'objets (réponse de IT commencent après 100 ms). Ces résultats impliquent de nouvelles contraintes très fortes sur les traitements visuels, et posent de nombreuses questions sur les mécanismes mis en jeu.

Mes travaux de thèse ont donc consisté à réaliser une série d'expériences utilisant ce protocole de choix saccadique afin de mieux comprendre les mécanismes impliqués dans cette tâche, ainsi que la nature des informations qui pouvaient être utilisées en un temps si court. Dans un premier temps, nous verrons que ce protocole permet de mettre en évidence des effets invisibles avec l'ancienne tâche basée sur une réponse manuelle. L'apparition de ces nouveaux phénomènes pourrait s'expliquer par le fait qu'en allant chercher l'information plus tôt, on aurait accès à une fenêtre temporelle plus précoce et donc un état de l'information plus *brut* (section 2.2). Une deuxième étude questionnera directement la nature de l'information sensorielle utilisée par le système visuel pour produire des réponses si rapides (section 2.3). Cet accès privilégié à une fenêtre temporelle précoce sera ensuite utilisé pour préciser deux effets bien établis par les études en go/no-go : l'influence du traitement du contexte sur la reconnaissance d'objet, et l'accès aux différents niveaux de catégorisation (chapitre 3). Dans le chapitre 4, je testerai explicitement l'hypothèse de différence de latence proposée initialement dans l'article 1 en proposant un modèle de décision sensorielle. Finalement, je synthétiserai dans le dernier chapitre les idées proposées au cours de ce travail.

Chapitre 2

Traitements visuels précoces et détection de visages

Sommaire

La détection de visages consiste à détecter, puis à localiser un visage dans le champ visuel. Les recherches en vision par ordinateur se sont depuis longtemps attelées à créer des algorithmes puissants permettant cette opération. Ceux-ci sont maintenant implémentés dans toutes sortes de périphériques grand public (appareils photo et caméras numériques, webcams, téléphones portables). La plupart de ces algorithmes sont de purs produits de l'ingénierie et du traitement d'images. De façon surprenante, cette question a eu beaucoup moins de succès auprès des neurosciences de la vision. Celles-ci se sont en effet principalement focalisées sur les processus de reconnaissance de visages, comme par exemple la reconnaissance des visages au niveau individuel, ou plus simplement la discrimination de visages entre eux. Ce champ de recherche a mené à des

modèles extrêmement développés (voir Tsao et Livingstone, 2008, pour une revue de la littérature). Plusieurs aires cérébrales seraient dédiées à leur traitement, les deux aires principales mises en évidence étant l'OFA et la FFA (Rossion et Gauthier, 2002; Kanwisher, 2000) grâce à des protocoles de soustraction en IRMf . Électrophysiologiquement, le traitement spécifique des visages, par contraste aux autres objets, se manifesterait notamment par une plus grande amplitude de l'onde N170 (apparaissant 170 ms après l'affichage des stimuli, Rossion *et al.* 2000; Rossion et Gauthier 2002, voir l'introduction de l'article 1 pour plus de détails et de références sur les données électrophysiologiques). Cependant, la détection de visages par le système visuel, en dépit de son apparente plus grande simplicité, n'en reste pas moins un pré-requis indispensable avant tout traitement plus avancé, telle que l'identification ou la reconnaissance d'expression. Comme nous le verrons par la suite, la simplicité de cette tâche va permettre au cerveau de s'imposer un autre défi : la rendre quasiment immédiate.

Du fait de la littérature limitée concernant la détection de visage dans le domaine des neurosciences, je présenterai aussi brièvement les travaux réalisés dans le domaine de la vision par ordinateur. Nous verrons ensuite une première série d'études expérimentales réalisées durant cette thèse qui ont permis de démontrer la spécificité des résultats qui pouvaient être obtenus grâce à la tâche de choix saccadique (section 2.2). Notamment, nous avons démontré la possibilité d'initier des saccades sélectives vers les visages beaucoup plus rapidement que ce qui avait été observé jusqu'ici, ces saccades rapides apparaissant avant même les premières activations sélectives des aires de reconnaissance haut-niveau comme le cortex inférotemporal. Ces résultats suggèrent donc fortement que l'information nécessaire à la réalisation de la tâche pourrait être prise en amont dans le système visuel. La deuxième série d'études qui clôturera ce chapitre a consisté à étudier une hypothèse spécifique concernant les informations bas-niveau qui pourraient être le support de ces réponses rapides (section 2.3).

2.1 Etat de l'art en détection de visage

La détection et la reconnaissance de visages semblent être deux processus dissociés par le système visuel. Une première raison est qu'ils reposent sur des exigences opposées : l'identification précise d'un visage requiert une analyse précise pour différencier les exemplaires entre eux, alors que la détection doit se baser sur un dénominateur commun à tous ces exemplaires. L'étude de patients prosopagnosiques, une forme particulière d'agnosie les rendant incapables de reconnaître les visages en général, a permis de mettre en évidence cette dissociation. En effet, la plupart de ces patients, malgré leurs résultats médiocres en reconnaissance, conservaient des performances tout à fait respectables en détection (Tsao et Livingstone, 2008). Cependant, la plupart de ces études ont été faites avec des protocoles de catégorisation qui pourraient être inadap-

tés (Garrido *et al.*, 2008). Dans ces expériences, les sujets devaient dire si l'image qui leur était présentée était un visage ou non, les images non-visage étant des visages "mélangés", c'est à dire qu'elles étaient composées de tous les éléments locaux d'un visage mais dans le mauvais ordre (par exemple la bouche au niveau du front, le nez au niveau du menton, etc.). Ce type de procédé a aussi amené d'autres types de conclusions, notamment chez le sujet sain (voir Lewis et Ellis, 2003, pour une revue de ces expériences), comme l'absence d'effet *pop-out* d'un visage dans une tâche de recherche visuelle (Nothdurft 1993 avec des visages dessinés, Brown *et al.* 1997 avec des images naturelles).

Malheureusement, une tâche de catégorisation visage/non-visage avec ce type de distracteur reste très éloignée de ce en quoi consiste réellement la détection de visages dans un environnement naturel. Ainsi, elle se déroule rarement en présence de visages mélangés autour, mais consiste plutôt à détecter la présence d'un visage au sein d'une scène riche et complexe. C'est donc plutôt dans ce type de tâche que se trouve la clé des mécanismes de détection de visages. Suivant cette idée, une série d'expériences a été menée par Lewis et Edmonds dans lequelles ils ont ré-évalué les conclusions sur l'effet *pop-out* des visages à l'aide d'un protocole plus réaliste. Pour ceci, ils découpaient une scène en plusieurs carrés, l'un d'eux contenant le visage. La moitié des scènes présentées contenaient effectivement un visage dans un des carrés, et ils faisaient varier le nombre de carrés distracteurs. Ceci revenait donc à un protocole classique de recherche visuelle, mais où les distracteurs correspondaient à ce qui est proposé généralement au système visuel dans le monde réel (il est important de souligner que les différents carrés étaient mélangés, pour casser la structure globale de la scène). Les conclusions se sont avérées très différentes des précédentes avec un effet *pop-out* très clair, ainsi qu'un avantage pour les scènes non-mélangées sur les scènes mélangées, l'organisation spatiale globale aidant donc aussi la détection (Lewis et Edmonds, 2003). Par la suite, ces mêmes auteurs ont affiné ces résultats et ont montré sa robustesse à de nombreuses modifications des images (couleur, flou), seule l'inversion de la polarisation semblant pouvoir annuler cet effet (Lewis et Edmonds, 2005). Une autre étude a confirmé l'effet de *pop-out* pour les visages (Hershler et Hochstein, 2005), faisant naître un débat sur la nature précise de cet effet. VanRullen a ainsi proposé que l'effet *pop-out* serait uniquement guidé par le spectre d'amplitude spécifique des visages dans l'espace fréquentiel de Fourier (voir VanRullen 2006, puis la réponse de Hershler et Hochstein 2006). En effet, et comme il sera aussi question dans ce chapitre (section 2.3), certaines propriétés des images de visage dans cet espace pourraient attirer le regard/l'attention automatiquement (Honey *et al.*, 2008).

Dans un autre domaine, de nombreuses recherches en vision par ordinateur se sont intéressées au problème de la détection de visages. Le but étant souvent de créer des

algorithmes de reconnaissance faciale qui doivent trouver des applications concrètes, les chercheurs de ce champ ont dû prendre en compte cette étape préliminaire. Un grand nombre de méthodes différentes ont été utilisées, certaines faisant directement écho à des processus biologiquement plausibles. La plupart de ces méthodes fonctionnent cependant sur un base commune aux méthodes de classification : l'algorithme commence par extraire certaines caractéristiques d'une base d'images d'entraînement, puis scanne chaque nouvelle image pour retrouver les caractéristiques apprises à différentes positions et échelles. Les caractéristiques utilisées peuvent être basées sur les pixels, sur l'extraction de bords locaux, ou encore sur un traitement en ondelettes de Haar (voir Viola et Jones 2004 pour le modèle le plus utilisé aujourd'hui inspiré par les ondelettes de Haar, Yang 2009 et Yang *et al.* 2002 pour des revues). L'analyse multi-échelle est certainement aussi la méthode utilisée par le système visuel pour ces tâches de détection. La question est donc de savoir quelles caractéristiques (locales et/ou globales) sont utilisées plus précisément. Une meilleure compréhension des mécanismes mis en jeu par le système visuel humain peut s'avérer décisive pour répondre à cette question.

2.2 Détection ultra-rapide de visages

2.2.1 Résumé de l'étude

Les travaux précédents de l'équipe ont permis de montrer que le système visuel humain était extrêmement rapide et performant pour détecter la présence d'un animal dans une scène naturelle complexe. En particulier, en utilisant une tâche de choix saccadique, Holle Kirchner et Simon Thorpe (2006) ont montré que lorsque deux images sont flashées simultanément sur la gauche et la droite d'un écran, les sujets peuvent initier des saccades du côté de l'image contenant un animal dès 120-130 ms après l'affichage de l'image. Dans l'étude présentée ici, nous montrons que si la cible est un visage humain, les saccades peuvent être initiées encore plus rapidement, avec des temps de réaction pour les saccades sélectives les plus rapides à 100-110 ms (temps de réaction moyen de 140 ms). Il semble que ces saccades ultra-rapides ne soient pas complètement sous le contrôle des sujets. En effet, lorsque la cible est un véhicule (avec un visage en distracteur), les sujets continuent à être biaisés vers le côté du visage pour les saccades les plus rapides. Pour terminer, nous avons aussi testé la possibilité pour les sujets d'initier ces saccades rapides lorsque les images n'étaient pas présentées à gauche et à droite, mais en haut et en bas. De tels résultats imposent de très fortes contraintes sur les traitements requis pour la détection et la localisation de visages.

2.2.2 Article 1 : Fast saccade towards faces : face detection in 100 ms

Fast saccades toward faces: Face detection in just 100 ms

Sébastien M. Crouzet Centre de Recherche Cerveau and Cognition, UMR, CNRS, Université Toulouse, Toulouse, France

Holle Kirchner Centre de Recherche Cerveau and Cognition, UMR, CNRS, Université Toulouse, Toulouse, France

Simon J. Thorpe Centre de Recherche Cerveau and Cognition, UMR, CNRS, Université Toulouse, Toulouse, France

Previous work has demonstrated that the human visual system can detect animals in complex natural scenes very efficiently and rapidly. In particular, using a saccadic choice task, H. Kirchner and S. J. Thorpe (2006) found that when two images are simultaneously flashed in the left and right visual fields, saccades toward the side with an animal can be initiated in as little as 120–130 ms. Here we show that saccades toward human faces are even faster, with the earliest reliable saccades occurring in just 100–110 ms, and mean reaction times of roughly 140 ms. Intriguingly, it appears that these very fast saccades are not completely under instructional control, because when faces were paired with photographs of vehicles, fast saccades were still biased toward faces even when the subject was targeting vehicles. Finally, we tested whether these very fast saccades might only occur in the simple case where the images are presented left and right of fixation by showing they also occur when the images are presented above and below fixation. Such results impose very serious constraints on the sorts of processing model that can be invoked and demonstrate that face-selective behavioral responses can be generated extremely rapidly.

Keywords: fast visual processing, face detection, visual processing time, saccade, eye movement

Citation: Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision, 10*(4):16, 1–17, http://journalofvision.org/10/4/16/, doi:10.1167/10.4.16.

Introduction

Measurements of processing speed in the visual system can be very useful for constraining models. For example, in a manual go/no-go task, subjects can reliably release a button when an animal is present in a natural scene from around 300 ms after stimulus onset (although mean reaction times are longer), and in the same situation, there is a differential EEG response between target and distractor trials that appears only 150 ms after stimulus onset (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; Thorpe, Fize, & Marlot, 1996). These latencies are undoubtedly very short given the computational complexity of the task and have led to the suggestion that at least some sorts of high-level visual tasks can be performed on the basis of a single feed-forward sweep through the visual system (Serre, Oliva, & Poggio, 2007; Thorpe & Imbert, 1989; VanRullen & Thorpe, 2002). Nevertheless, there is evidence that processing involving feedback can occur very rapidly. For example, the question of whether a particular region of the image is foreground or background affects activity in areas such as V1 within a few tens of milliseconds of the start of the neural response (Qiu, Sugihara, & von der Heydt, 2007; Roelfsema, Tolboom, & Khayat, 2007). As a consequence, even processing involving both a feed-forward and a feed-back pass could be possible very rapidly (Epshtein, Lifshitz, & Ullman, 2008). For this reason, precise measurements of processing time are likely to become even more important for distinguishing between processing that can be achieved with a single feed-forward pass and processing that leaves enough time for both bottom-up and top-down mechanisms to be involved.

While much of the evidence for ultra-rapid scene processing has come from using a manual go/no-go categorization task coupled with event-related potentials, it has become clear that there are limitations to this approach. One of the strongest arguments in favor of the idea that the differential EEG response at around 150 ms is indeed related to categorization comes from the fact that the differential activity can be modulated by changing the target category (VanRullen & Thorpe, 2001b). Thus, when subjects were required to switch the target category from "animal" to "means of transport" in different blocks, there were differences in the ERP response from around 150 ms that depended not on the physical characteristics of the stimulus but rather on the status ("target vs. distractor") of the image. This effectively rules out the possibility that the differential activity simply results from

doi: 10.1167/10.4.16 Received November 26, 2009; published April 28, 2010 ISSN 1534-7362

irrelevant low-level differences between the stimuli. However, other studies pointed out that a considerable part of the differential activity occurring at short latencies was not affected by the status of the stimulus (Johnson & Olshausen, 2003, 2005). This phenomenon was particularly marked in a study using human and animal faces as stimuli. Subjects were extremely good at responding selectively to either human or animal faces and could switch virtually effortlessly between the two different target categories from block to block (Rousselet, Mace, & Fabre-Thorpe, 2003), achieving accuracy levels of over 97% correct, irrespective of whether the target category was human or animal. Despite this, the analysis of the simultaneously recorded ERP signals failed to find any evidence that short latency responses (i.e., at latencies below about 180 ms) could be modulated by whether the particular stimulus was a target or not (Rousselet, Mace, Thorpe, & Fabre-Thorpe, 2007). There were strong differences in the ERP signals recorded with human and animal faces, but those differences were unaffected by whether the subject was treating the image as a target or not. If a dependence on task status is needed to be able to infer that a particular neural response is related to high-level processing, then it would be natural to conclude that no influence of such high-level factors is visible until around 180 ms in this task.

Does this mean that the differential brain responses seen earlier should be dismissed as simple artifacts due to irrelevant low differences between images? One type of result that argues strongly against this are recent studies using a saccadic choice task that show that useful behavioral responses can be generated well before the 150–180 ms latency value suggested by the task-dependent ERP effects. In the first such study, Kirchner and Thorpe reported that when two natural scenes are simultaneously flashed left and right of fixation, reliable saccades to images containing animals can be initiated as early as 120–130 ms after images onset (Kirchner & Thorpe, 2006). Given that motor preparation presumably needs at least 20 ms, this implies that the underlying visual processing may need only 100 ms, considerably earlier than the 150-ms latency of the first differential activity.

In the present study, we have used a similar saccadic choice task protocol to the one used previously by Kirchner and Thorpe, but with another type of highly significant visual stimulus—photographs of human faces. Experiment 1 directly compared saccadic reaction times for three stimulus categories—animals, faces, and vehicles—and demonstrated a very impressive result for faces for which the fastest reliable saccades were seen at latencies as early as 100–110 ms. Then, Experiment 2 revealed a search asymmetry between faces and vehicles: subjects found it much easier to saccade toward the faces than toward the vehicles. Interestingly, they were only able to fully overcome this bias for saccades with relatively long latencies. Finally, Experiment 3 allowed us to test a simplistic model of the task based on a comparison between activation in the left and right hemispheres. It showed that the fast detection of faces is not restricted to a horizontal display arrangement and must presumably rely on a more complex process than a simple comparison between activation levels in the two hemispheres.

The results, some of which have been presented previously (Crouzet, Kirchner, & Thorpe, 2008), fit with a large range of previous studies that have demonstrated that faces have a special status. More importantly, they impose some very serious temporal constraints on the underlying processing. Specifically, face-selective behavior can be initiated at latencies so short that there may not be enough time to complete the first wave of processing in the cortical areas of the ventral stream. If so, other possibilities including the involvement of subcortical pathways may be involved.

Experiment 1: Comparison between three target categories

Experiment 1 examined whether performance in the Saccadic Choice Task varies when different object categories are used as the target (faces, animals, or vehicles). Each of these target categories was displayed in combination with the same set of "neutral distractors", corresponding to various natural scenes. Thus, any differences observed between the three conditions will reflect difference of processing time between the three categories of objects, rather than differences between the distractor stimuli.

Methods

Participants

Eight volunteers (7 men; mean age 24.5 years, ranging from 22 to 31 years) with normal or corrected-to-normal vision participated in a 2-AFC saccadic choice task. They all gave written informed consent to participate in the experiment.

Stimuli

One thousand photographs selected from the Corel Photolibrary database or downloaded from the Internet were used to set up four object categories of 250 natural scenes: faces, animals, vehicles, and neutral distractors. The neutral distractor category was composed of a range of images that all contained a salient object in the foreground. All the images were converted to grayscale and resized to 330 × 330 pixels. The global contrast of each image was reduced to 80% of the original image,

Figure 1. Examples of images used in this study.

allowing us to adjust the mean luminance of each image to a grayscale value of 128. (Guyonneau, Kirchner, & Thorpe, 2006). The complete set of images can be provided on request (Figure 1).

Apparatus

Participants viewed the stimuli in a dimly lit room with their heads on a chin rest to maintain the viewing distance at 60 cm. Stimuli were presented on an IIYAMA Vision Master PRO 454 monitor with the screen resolution set to 800 × 600 pixels and a refresh rate of 100 Hz. The centers of the two images were always 8.6° from the fixation cross, resulting in a retinal size for each image of 14° by 14°. The experiment was run using the software Presentation 9.9 (Neurobehavioral Systems).

Protocol

The experiment was performed using a Saccadic Choice Task, a similar protocol to the one used by our team in previous studies (Guyonneau et al., 2006; Kirchner & Thorpe, 2006) with the exception that the natural scenes were displayed during 400 ms rather than flashed for 20 ms. The original reason for using such short presentation times when using a conventional manual go/no-go protocol was to exclude the possibility of ocular exploration (Thorpe et al., 1996). However, in the present experiments, we specifically required the subjects to make saccades. Since we were recording the eye movements, the original argument for using flashed presentations was now obsolete. Preliminary experiments had already shown that, contrary to what might have been expected, this longer presentation durations significantly shortened the mean RT. Specifically, it appears that using longer presentation times reduces the number of late saccades, resulting in a leftward shift and a sharpening of the Saccadic Reaction Time (SRT) distribution. An explanation could be that the offset of the images in the original protocol perturbed the initiation of some saccades, resulting in a right-biased distribution. A second difference with respect to the previous studies was that the background screen was set at a grayscale value of 128 rather than black.

Observers had to keep their eyes on a black fixation cross that disappeared after a pseudo-random time interval (800–1600 ms), leaving a 200-ms time gap before the presentation of the images (Fischer & Weber, 1993; Kirchner & Thorpe, 2006). The use of such a gap allows saccades to be initiated more rapidly. Two natural scenes,

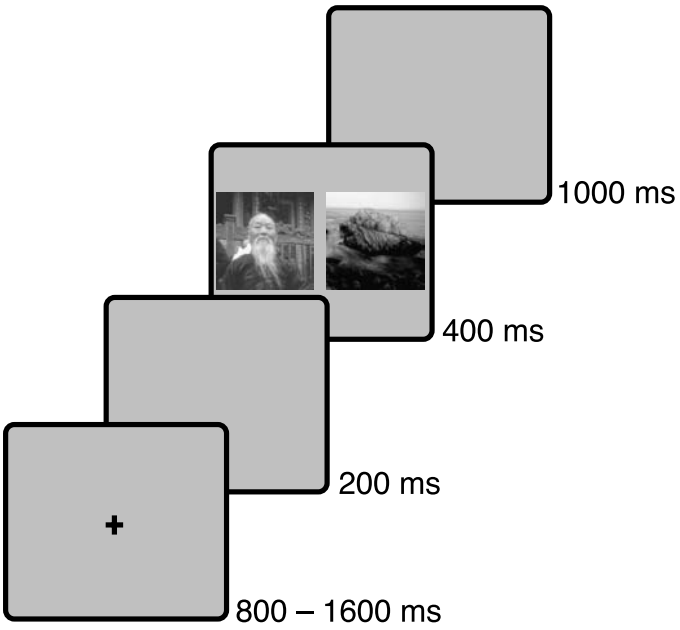

Figure 2. Protocol: The saccadic choice task. Observers had to fixate a cross in the center during a pseudo-random time (800–1600 ms). After a gap of 200 ms, 2 images were displayed left and right of fixation for 400 ms. Observers then had 1000 ms to prepare for the next trial.

one target and one distractor, were then displayed on each side of the screen for 400 ms (see Figure 2). The task was to make a saccade as quickly and as accurately as possible to the side where an object belonging to the target category has appeared.

Using a within-subject design, three object categories were tested here (faces, animals, and vehicles). Each subject saw each target image once during the experiment and thus performed 250 trials in each condition, divided into blocks of 50 trials. The "neutral distractor" images were the same in the three conditions, so each one was seen several times by each participant. The order of the three conditions was counterbalanced across participants. Each block was preceded by a training session of 50 trials with images not used in the experiment.

Response recording and detection

Eye position was recorded using horizontal EOG electrodes (1 kHz, low pass at 90 Hz, notch at 50 Hz, baseline correction [−400:0] ms; NuAmps, Neuroscan). Saccadic Reaction Time (SRT) was determined offline as the time difference between the onset of the images and the start of the saccade. Each trial was verified by the experimenter to make sure that only the largest inflection (if any) was taken as a real saccade (see Kirchner & Thorpe, 2006 for more detailed information about the procedure); 15.2% of trials (912 of 6,000) had to be excluded because of a noisy eye signal, but this percentage was evenly spread across conditions (face task = 16%, animal task = 14.6%, vehicle task = 15%).

Minimum reaction times

To determine a value for the minimum SRT, we divided the saccade latency distribution of each condition into 10-ms time bins (e.g., the 120-ms bin contained latencies from 115 to 124 ms) and searched for bins containing significantly more correct than erroneous responses using a χ^2 test with a criterion of $p < 0.05$. If 5 consecutive bins reached this criterion, the first was considered to correspond to the minimum reaction time.

Results and discussion

The principal finding from this first study was that subjects were fast and accurate in all three conditions (see

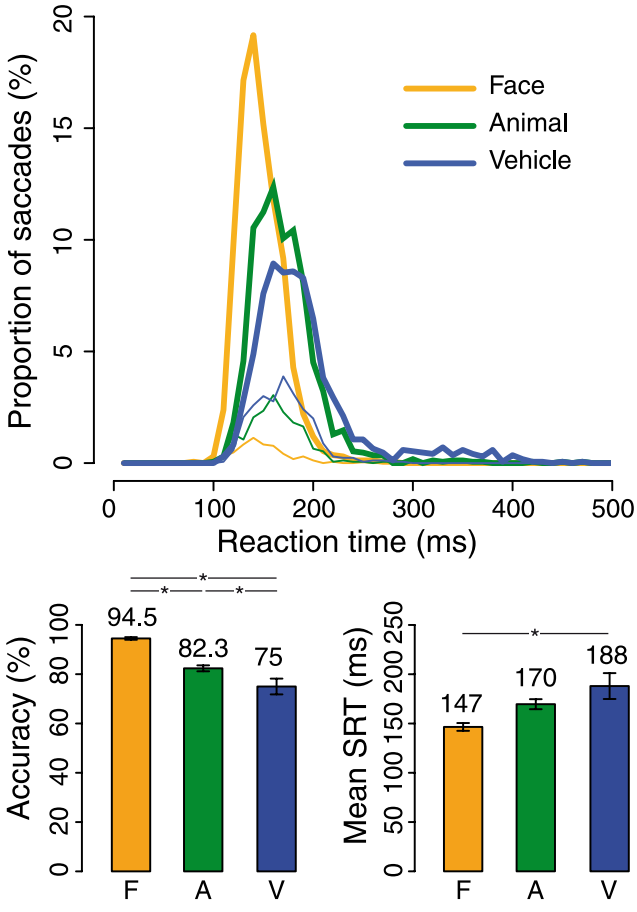

Figure 3. Experiment 1. (Top) Distributions of SRT for 3 different target categories: face, animal, vehicle. Correct responses are shown in thick lines, incorrect as thin lines. (Bottom) Mean accuracy and SRT in the 3 conditions. Errors bars are *SEM*.

Figure 3). Ultra-rapid processing of objects in the saccadic choice task is clearly not restricted to animals but can also be extended to vehicles and human faces. However, and contrary to what has been shown using manual response (Rousselet et al., 2003; VanRullen & Thorpe, 2001a), our results showed a clear ordering between categories for both SRT and accuracy. A one-factor ANOVA analysis showed that there was a global effect of the category used as target on both mean SRT ($F(2,14) = 10.622$, $p < 0.01$) and accuracy ($F(2,14) = 26.031$, $p < 0.001$). A post-hoc Tukey analysis for multiple comparisons showed a progressive increase on accuracy from the "vehicle" condition (75%) to the "animal" (82.4%) and "face" (94.5%) conditions. Mean SRT was only significantly different between the face (147 ms) and vehicle (188 ms) conditions.

Different processing times for different object categories

In addition to the very clear differences in mean reaction times seen for the three object categories, there were also very striking differences in the minimum reaction time values (i.e., the first bin of at least 5 consecutive bins in the reaction time distribution where there was a significantly higher proportion of correct than erroneous responses). In the case where the target category was "animal", the value obtained replicated Kirchner and Thorpe's (2006) study with a minimum SRT of 120 ms. Interestingly, this minimum SRT was clearly higher for vehicles (140 ms) and lower for faces (110 ms). Together, the results demonstrate a very clear advantage for the processing of faces over animals and vehicles when observers had to discriminate these object categories from "neutral distractors".

Experiment 2: Faces vs. vehicles

Previous studies using a manual go/no-go protocol had shown that subjects can switch from one target category to another in different blocks with little cost in terms of either accuracy or reaction time. This was seen for both the situation where the target categories are animals and means of transport (VanRullen & Thorpe, 2001a), as well as with humans versus animals (Rousselet et al., 2003). In Experiment 2, we ask whether this ability to switch between target categories also exists in the case of the saccadic choice task. Specifically, we designed an experiment in which the subjects had to discriminate directly between two object categories: faces and vehicles. This design allowed us to directly compare processing times between these two categories of objects, but additionally, we can see if the task can be reversed under voluntary control. As an example, if a subject was instructed to treat faces as targets and vehicles as distractors in a first block, subsequent blocks could require the reverse configuration, with vehicles as targets and faces as distractors. The results reveal a clear asymmetry, with saccading to faces being considerably faster and more accurate than to vehicles. Indeed, subjects had great difficulty in making fast saccades toward vehicle targets.

Methods

Participants

Eight volunteers (5 men; mean age 26.9 years, ranging from 23 to 34 years) with normal or corrected-to-normal vision participated in a 2-AFC saccadic choice task. They all gave written informed consent to participate in the experiment.

Stimuli

In order to have a more controlled set of stimuli, 200 photographs were selected from the Corel database and the Internet to generate two object categories, each with 100 images: faces and vehicles. The faces were 50% men and 50% women, while the vehicles were 50% cars and 50% trains. Each subcategory was divided equally into close-up and mid-distance views. Manipulations on the luminance and contrast of each image were the same as in Experiment 1.

Apparatus and protocol

The design was unchanged from Experiment 1 with the following exception: each subject saw each image four times, both as target and distractor and in both the left and right hemifields. Specifically, each participant performed 200 trials in each of the two conditions (face and vehicle). The order of the two conditions was counterbalanced across participants; 23% of trials (746 of 3,200) had to be excluded because of a noisy eye signal, but this percentage was evenly spread across conditions (face task = 22.3%, vehicle task = 24.4%).

Results and discussion

The first main result of Experiment 2 is that even with another object category as distractor, here vehicles, saccades toward faces can still be initiated very rapidly. Overall accuracy was 89.6%, with a mean SRT of 138 ms. Remarkably, the earliest reliable saccades appeared just 100–110 ms after scene onset. However, the situation was markedly different when the target category was vehicle. In this case, the values for mean SRT (167 ms) and accuracy (71%) were both significantly poorer than with faces ($F(1,7) = 84.723$, $p < 0.001$ and $F(1,7) = 44.867$, $p < 0.001$, respectively; Figure 4).

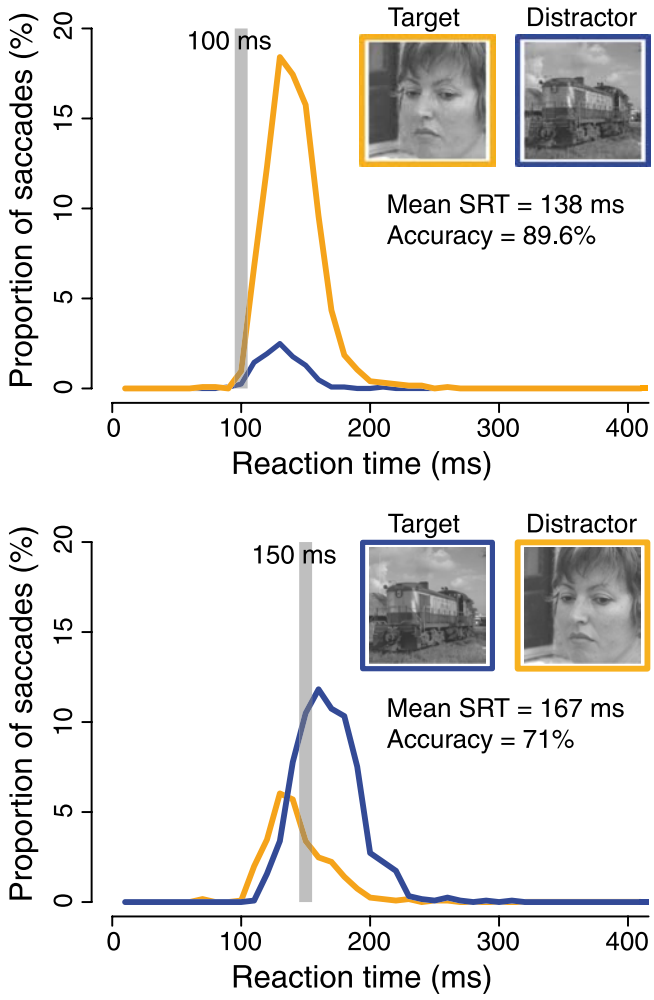

Figure 4. (Top) Distribution of SRT over all subjects when the task is to saccade toward faces (responses toward faces in orange, vehicles in blue). (Bottom) Distribution of SRT for all subjects when the task is to saccade toward vehicles. The gray vertical bar indicates the bin where correct responses start to significantly outnumber errors.

Even more striking was the distribution of response in the 100–140 ms time window. There was a tendency of saccades initiated in this time range to go toward the side with the faces, even if the task is to go to vehicles. Thus, saccades initiated before 140 ms seemed to be hard to control.

An interesting observation is that if we divide the data according to the position of the target (left or right, Figure 5), there is a clear tendency for subjects to be faster and more accurate when the target is on the left. This tendency can be observed in the face task (left: 135 ms and 95.8%; right: 143 ms and 83.3%) and in the vehicle task (left: 165 ms and 77.2%; right: 170 ms and 64.8%) and is significant for mean RT ($F(1,7) = 6.4953$, $p < 0.05$) although not for accuracy ($F(1,7) = 4.0321$, $p = 0.084$). Thus, if the target is in the left hemifield, participants produced fewer errors and their correct responses had a shorter mean SRT. When the target was in the right hemifield, participants made more early errors. Furthermore, many of the errors in the face task were made when the target is on the right and the distractor on the left, especially on fast saccades. For example, the specific pattern observed in Figure 4 with more responses toward faces than vehicles in the 100–140 ms time window is largely the result of the situation when the face is on the left. A similar tendency to produce more saccades on the left has also been reported when people look at chimeric faces (Butler et al., 2005). These left hemifield biases could be related to the well-known fact that neural responses in the right hemisphere are reliably stronger than on the left, a result that has been repeatedly seen in both fMRI studies (e.g., Hemond, Kanwisher, & Op de Beeck, 2007; Kanwisher, 2000) and ERPs (e.g., Jacques & Rossion, 2009; Rousselet, Mace, & Fabre-Thorpe, 2004). Indeed, the existence of a left hemifield advantage when saccading to faces supports the hypothesis that the saccadic choice task really does involve face processing mechanisms. If it were simply a bias toward making saccades toward the left, the same biases would be expected with any sort of target.

Additionally, half the stimuli were close-up views (CV), and the other half mid-distance views (MV). This allowed a post-hoc analysis of the effect of object size to be performed. It showed that there was absolutely no effect of the size of the target face on either SRT or on accuracy. This effect is null for the "face" (CV: 168 ms and 70%; MV: 165 ms and 72%) as well as for the "vehicle" conditions (CV: 138 ms and 90%; MV: 137 ms and 89%). Further studies to look at the effect of object size on discrimination performance in this task in a more systematic way would be of considerable interest.

In summary, performance in this task was remarkably fast and efficient. Subjects were able to move their eyes selectively to the side where the scene contains a face category target in as little as 100–110 ms. It is clear that we would not argue from this result that a face can be "recognized" in just 100 ms. Nevertheless, the result demonstrates that the visual system needs only around 100 ms to initiate an eye movement toward a face. Furthermore, the fact that subjects had such difficulty in reversing the task and saccading toward the vehicle suggests that this attractivity might be effectively hard-wired.

Experiment 3: Horizontal versus vertical positioning

One factor that might contribute to this remarkable level of performance may lie in the design of the protocol, in which the two images are displayed to the left and right of fixation. As a result, the two images will effectively be

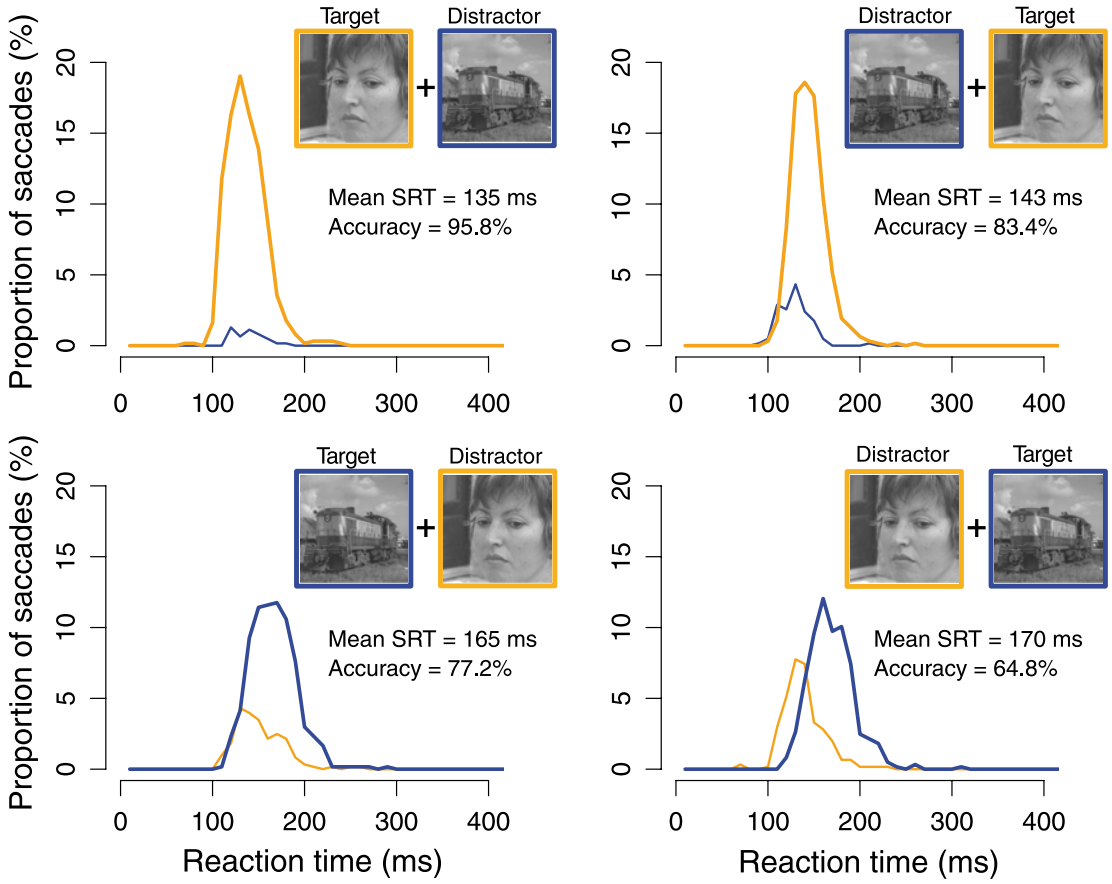

Figure 5. Distributions of SRT over all subjects when the task is to saccade toward faces (top row) or vehicles (bottom row) and when the target is on the left (left column) or on the right (right column). Correct responses are in thick lines, incorrect are in thin lines.

processed separately by the two hemispheres (at least initially) and this may provide a situation that is particularly favorable. Potentially, the task could be performed by comparing the activation in the two halves of the brain, and initiating the saccade to the side that has the strongest (or earliest) activation. This hypothesis was tested in Experiment 3 in which subjects were asked to perform the same task with either the images displayed horizontally (to the left and right of fixation) or vertically (above and below the fixation point). Experiment 3 also examined the difference between a saccadic choice task in which two images are presented at the same time and subjects required to saccade to a target, and simple detection task in which only one image is presented on each trial. As in Experiment 2, subjects were required to perform the task with faces and vehicles images, varying the target category in different blocks.

Methods

Participants

Four volunteers (2 men; mean age 32.7 years, ranging from 23 to 50 years) with normal or corrected-to-normal vision participated in a 2-AFC saccadic choice task. They all gave written informed consent to participate in the experiment.

Stimuli

Unchanged from Experiment 2 (Figure 6).

Protocol

The protocol was unchanged from Experiments 1 and 2 with the following exceptions. The experiment was divided in two sessions, each comprising 12 blocks of 50 trials. The first two blocks and the last two blocks in each session used a Simple Detection Task with only one image on each trial and no distractor, where the two categories of target (faces and vehicles) were mixed in the same block. Within each two-block group, one block used the images in the horizontal arrangement, while the other used a vertical one. For this Simple Detection Task, subjects were instructed to saccade as fast as possible on the side where there was an image, independently of the category of the image. It is important to notice that in the case of this Detection Task, no classification was needed.

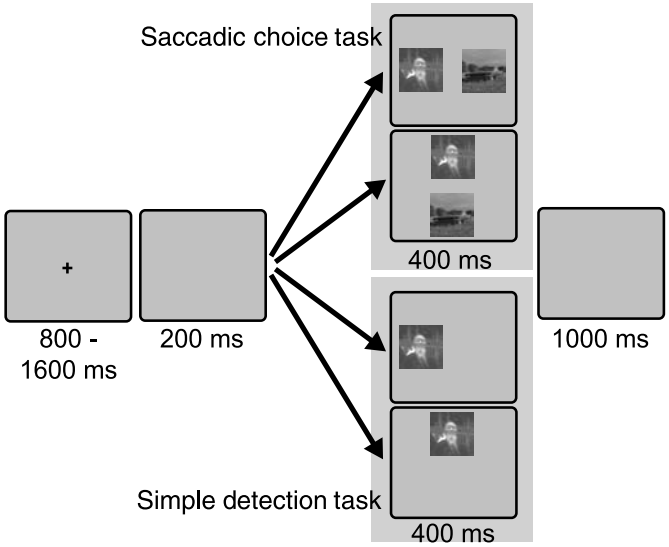

Figure 6. Design of Experiment 3. The protocol was similar to the one used in Experiments 1 and 2. After the 200-ms gap, and following a block design, participants had to perform a task in which either one image was presented (two screens on the bottom of the figure—Simple Saccadic Detection Task) or two images are presented simultaneously (two screens on the top-Saccadic Choice Task). In both cases, the images can be displayed horizontally or vertically.

In the middle part of each session, the subjects performed the Saccadic Choice Task with either four blocks with Faces as targets followed by four blocks with Vehicles as targets, or the contrary. In addition, the arrangement of the stimuli was varied with two blocks of vertically arranged stimuli alternating with two blocks using the horizontal arrangement. All the different orders were counterbalanced across the four subjects and the two sessions. Image size was still set at 330 × 330 but the resolution of the screen was increased to 1024 × 768 to allow the images to be displayed vertically. As a consequence, the retinal size of the images was 11° by 11°, and the center of images was 6.8° from the center of the screen.

Eye movement recording

Unlike Experiments 1 and 2, in this experiment the eye movements were monitored using a Chronos Eye Tracker (Chronos Vision, Berlin, Germany). This infrared tracking system samples eye position at 200 Hz binocularly. Saccade detection was performed offline, based on a velocity criterion and all the saccades were verified by the experimenter. Only the first saccades to end beyond 4° of eccentricity (corresponding to a minimum of 25% of the width of the image) were included in the analysis; 10.9% of trials had to be excluded using this criterion or because of a poor detection of the pupil. Before each block, an 8-point calibration procedure was performed.

Results and discussion

Vertical vs. horizontal display

As can be seen from Table 1, performance when the images were arranged vertically remained very good and was quite similar to the results obtained with the horizontal arrangement. Indeed, there was no overall effect of the arrangement of the stimuli (horizontal or vertical) on accuracy (as demonstrated by a two-way ANOVA). In contrast, the mean RTs for horizontal saccades were significantly shorter than vertical ones (167 ms and 178 ms, respectively; $F(1,9) = 9.1069$, $p < 0.05$). As might be expected from the results of Experiments 1 and 2, the nature of the target (face or vehicle) still has a strong

		Saccadic choice task			Simple detection task	
Target category	Target location	Mean SRT (ms)	Accuracy (%)	Min. SRT (ms)	Mean SRT (ms)	Min. SRT (ms)
Face	Left	150 ± 14	95.5 ± 2		119 ± 12	
	Right	159 ± 13	84.3 ± 5.9		133 ± 14	
	Horizontal display	154 ± 13	89.8 ± 3	100	126 ± 12	80
	Bottom	165 ± 17	85.4 ± 6.5		146 ± 11	
	Top	168 ± 11	86.6 ± 5.8		135 ± 10	
	Vertical display	166 ± 14	86.3 ± 4.2	110	139 ± 10	90
Vehicle	Left	176 ± 17	83 ± 9		125 ± 15	
	Right	185 ± 19	68.8 ± 9.1		147 ± 16	
	Horizontal display	180 ± 17	75.8 ± 6.8	170	136 ± 15	80
	Bottom	190 ± 20	67.1 ± 8		140 ± 13	
	Top	187 ± 13	74.3 ± 5.8		138 ± 10	
	Vertical display	189 ± 16	71 ± 4.7	190	139 ± 11	90

Table 1. Results for Experiment 3. Mean SRT and accuracy are presented for both the Saccadic Choice Task (with two simultaneously presented images) and the Simple Detection Task (a single image presented).

effect both on mean RT and accuracy ($F(1,9) = 48.7835$, $p < 0.001$; $F(1,9) = 31.5336$, $p < 0.001$, respectively). Thus, these results showed a slight difference between a horizontal and a vertical display.

However, a more detailed analysis that divided the results according to the precise location of the target (Left, Right, Top, Bottom) revealed that the difference between the results with horizontal and vertical displays was essentially due to the strong advantage when the target is presented on the left, an effect already seen in Experiment 2. Thus, it seems that in these experiments, saccading to the right, the top, or the bottom of the screen was roughly equivalent, but that saccading to the left was faster.

The critical point here was that subjects were still able to produce very fast responses in the saccadic choice task even when the stimuli were positioned vertically. This effectively rules out a simple "hemisphere comparison" hypothesis in which the eyes simply move because there is an imbalance of activity between the left and right hemispheres. This is because, presumably, when the images are positioned vertically the amount of activation in the two hemispheres will be roughly balanced. However, it is worth noting that there is also evidence that the processing of the upper and lower visual fields involves anatomically separate areas in extrastriate cortex. As a consequence, it may still be possible to envisage a competitive mechanism in which global activation levels in two separate brain structures are compared. Further experiments would be needed to test the limits of this sort of ability by, for example, presenting both the target and distractor stimuli in the same hemifield, or even in the same quadrant.

Saccadic choice task vs. simple saccadic detection task

A second major result emerging from Experiment 3 was the relatively small difference between SRT distributions in the simple saccadic detection task and the saccadic choice task, at least with faces as targets. The fact that there are essentially no errors in the simple saccadic detection task means that it is not useful to compare accuracy levels. Mean SRT values were 159 ms for faces versus 183 ms for vehicles in the saccadic choice task. The corresponding values were 131 ms and 137 ms in the simple saccadic detection task. A two-way ANOVA showed that there is still a significant effect of the category of the target ($F(1,9) = 22.9889$, $p < 0.001$), and a clear advantage for simple detection over the saccadic choice task ($F(1,9) = 144.6105$, $p < 0.001$). Furthermore, the interaction between the task and the category of the target is significant ($F(1,9) = 9.0315$, $p < 0.05$), meaning that the difference between saccadic choice and simple detection is much larger for vehicles than for face targets. This is also clear from looking at the minimum SRT because when the target was a face, the difference between minimum SRTs for the choice task and the simple detection task was only 20 ms. In comparison, when the target was a vehicle, this additional time cost was 80 ms.

General discussion

The present study used a saccadic choice task to investigate the time course of the processing involved in ultra-rapid detection of objects in natural scenes. The experiments follow on from an earlier study that had shown that subjects can make rapid and reliable saccades toward animal targets when two images are simultaneously flashed left and right of fixation (Kirchner & Thorpe, 2006). Experiment 1 showed that these very rapid saccadic responses can be initiated even more rapidly when human faces are the target category, with the fastest saccades being initiated from around 110 ms following the onset of the stimuli. Then, Experiment 2 showed that this strong bias toward saccading toward faces is very difficult to suppress, because even when subjects are actively trying to saccade toward vehicles, they still show a very clear tendency for fast saccades to be directed toward faces. Finally, Experiment 3 showed that this ability to initiate very fast saccades toward face targets is not restricted to the specific left/right design and thus cannot be explained by a simple comparison between activation levels in the two hemispheres.

Ultra-rapid processing of faces

The values reported here for saccadic reaction times with faces are remarkably short. In Experiment 2, in which faces were paired with photographs of cars and trains, the mean onset latency for saccades toward faces was a mere 138 ms. Virtually all the saccades were initiated in under 200 ms, but despite this, accuracy was a very respectable 89.6%. By examining the distribution of correct and erroneous saccades in each 10-ms bin of the reaction time distribution, we were able to show that the minimum reaction time in this task was only 100–110 ms. Such values put very severe temporal constraints on the underlying visual processing, especially when one takes into account the fact that the latencies obtained from the EOG and eye tracker data used in this study include the time needed to initiate the eye movement. Most neurophysiologists would allow roughly 20 ms for the activation of the brain stem structures involved in oculomotor control and the muscles of the eye itself. If true, then it appears that information about the presence of a face in an image may be available as little as 80 ms after the onset of the image. Such values are considerably shorter than

previous behavioral estimates of processing times in the human visual system (Thorpe et al., 1996) and are even a lot shorter than the 150-ms differential ERP response between targets and distractors that has previously been used as a measure of processing time.

The mean SRT values seen here are substantially shorter than those reported in the previous study by Kirchner and Thorpe (2006) in which the median SRT when using photographs of animals as targets was 228 ms. However, there are a number of differences between the two experiments that could explain why the reaction times were so much shorter here. In Experiment 1, we used a very similar situation with photographs of animals paired with complex natural scenes as distractors and obtained a mean SRT value of 170 ms. It seems likely that some of this reduction in reaction time results from the fact that in the current experiments, the two images remain present for 400 ms, instead of simply being flashed for 20 ms, as in the original study. It could be that removing the stimulus before the saccade has been initiated in the original design might interfere with saccade initiation. In contrast, by leaving the images on for 400 ms, the subject is in the very natural situation of initiating saccades toward a stimulus that is still present when the eyes arrive at their destination. Given that the aim of the experiment is to obtain the shortest realistic measurement of the time required to initiate a saccade toward a target, there seems to be little reason to continue with the original design, which only seems to introduce additional variability in the reaction time distribution.

Comparison with previous eye movement studies

The tendency of humans to look preferentially at faces when exploring visual scenes was already clear in the classic studies of Buswell (1935) and Yarbus (1967), and a recent study showed that similar biases also exist in chimpanzees (Kano & Tomonaga, 2009). However, it is important to realize that several factors may be involved in producing such biases (Henderson, 2003). For example, there may be a tendency to fixate faces for longer than other less interesting parts of the image. The most well widely used models of gaze control, such as Itti and Koch's (2000) saliency model rely on local variations in relatively low-level factors such as color, orientation, and luminance. While such models can account for a substantial proportion of real-world gaze patterns in humans (Parkhurst, Law, & Niebur, 2002; Peters, Iyer, Itti, & Koch, 2005; Tatler, Baddeley, & Gilchrist, 2005), there are a number of studies showing that such models cannot be considered complete (Birmingham, Bischof, & Kingstone, 2008; Cerf, Harel, Einhäuser, & Koch, 2008). Indeed, by changing the task requirements, it is possible to override these low-level biases (Einhauser, Rutishauser, & Koch, 2008).

A separate question concerns the issue of whether the very first saccades that are generated in response to a scene can be directed to important objects such as faces, and if they can, at what latency. The study by Kirchner and Thorpe had already demonstrated this for animal targets, and a recent study showed that these rapid saccades toward animals can be quite accurate in terms of localization (Drewes, Trommershaeuser, & Gegenfurtner, 2009). Another recent study by Fletcher-Watson, Findlay, Leekam, and Benson (2008) extended the use of the saccadic choice task to the detection of humans. Observers viewed two images presented at the same time on the left and right of a screen, only one of which contained a human. They reported that participants tend to saccade more on the side with the human, and that this bias was indeed seen for the very first saccades. The saccades were grouped into bins of 50 ms, and not surprisingly, the first two bins had very few responses. However, in the bin from 100 to 149 ms, 90% of the saccades were oriented toward the image containing a face. One particular feature of that study was that they included a task where the subjects had to saccade spontaneously to one of two images, with no task to perform. The fact that even here subjects showed a strong tendency to look toward the side with the human suggests that there may well be a built-in bias toward looking at humans.

Other data supporting the idea that faces can be processed very efficiently comes from the extensive literature on attentional capture (Bindemann, Burton, Hooge, Jenkins, & de Haan, 2005; Langton, Law, Burton, & Schweinberger, 2008; Ro, Russell, & Lavie, 2001; Theeuwes & Van der Stigchel, 2006; Vuilleumier, 2000) as well as recent studies reporting pop-out in displays containing large numbers of elements (Hershler & Hochstein, 2005, 2006; although see also Brown, Huey, & Findlay, 1997; Vanrullen, 2006). Additionally, the bias toward faces has also been seen using an anti-saccade protocol, which demonstrated that subjects have difficulty in looking away when a face is present (Gilchrist & Proske, 2006). Together, all these results point toward a real behavioral advantage for faces in a wide range of situations.

Underlying brain mechanisms

These behavioral effects are also reflected at the level of brain mechanisms. Numerous studies have suggested that faces may have a special computational status that would allow them to be processed more efficiently and faster than other classes of objects (Farah, Wilson, Drain, & Tanaka, 1998; Haxby, Hoffman, & Gobbini, 2000; Kanwisher, 2000; but see Tarr & Gauthier, 2000). Faces are known to activate spatially adjacent but distinct brain regions in both humans and monkeys (Freiwald, Tsao, & Livingstone, 2009; Sergent, Ohta, & MacDonald, 1992; Tsao & Livingstone, 2008). This has been shown in a

range of techniques including PET, fMRI, and single unit recording (Freiwald et al., 2009; Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999; Kanwisher, McDermott, & Chun, 1997; Puce, Allison, Gore, & McCarthy, 1995). Nevertheless, there is still debate about the nature and degree of specialization (Cohen & Tong, 2001; Downing, Jiang, Shuman, & Kanwisher, 2001; Haxby et al., 2001).

The fact that the earliest reliable saccades toward faces can be seen as early as 100–110 ms after stimulus onset places particularly severe constraints on the underlying brain mechanisms. In the present context, it is particularly important to look at experimental evidence on the speed with which information about faces can be processed. Following the earliest reports of face-selective Event Related Potentials (Jeffreys, 1989), much attention has been paid to the N170 potential that seems to be particularly strongly associated with face processing (Bentin, Allison, Puce, Perez, & Mccarthy, 1996; McCarthy, Puce, Belger, & Allison, 1999; for a recent review, see Rossion & Jacques, 2008). However, it seems likely that the N170 occurs too late to be directly involved in triggering the fastest saccades reported here. Nevertheless, there have been repeated reports of face-selective electrophysiological responses occurring at even earlier latencies. For example, Liu, Harris, and Kanwisher (2002) reported face-selective MEG activation at latencies of around 100 ms, and selective ERP responses to emotional faces have been reported with latencies of around 120 ms (Eimer & Holmes, 2002). There have even been reports of face-selective repetition-related effects at even shorter latencies, sometimes as early as 45–80 ms (George, Jemel, Fiori, & Renault, 1997; Mouchetant-Rostaing & Giard, 2003; Mouchetant-Rostaing, Giard, Bentin, Aguera, & Pernier, 2000) or even 30–60 ms (Braeutigam, Bailey, & Swithenby, 2001), but it has been unclear whether these very rapid differential effects are really related to face perception. Clearly, in the light of the present behavioral responses, it may be appropriate to reconsider the significance of these very early phenomena.

Another important source of information about processing speed is the results of single-cell recording studies in awake primates that have shown that face-selective neuronal responses can be seen from around 100 ms although the fastest single unit responses can be as early as 70 ms (Oram & Perrett, 1992). It is also important to determine precisely when information about object identity can be read out from the activity of a population of cells. This issue was addressed in a recent study that examined the responses of populations of single neurons in monkey inferotemporal cortex and found that decisions about both object category and identity could be made on single trials from around 100 ms using a temporal window of only 12.5 ms in duration (Hung, Kreiman, Poggio, & Dicarlo, 2005).

In contrast to work in monkeys, there have been relatively few single unit recording studies in humans (though see, for example, the work that has been done on recordings from the medial temporal lobe in epileptic patients; Mormann et al., 2008). Most human data comes from ERP and MEG recordings that are less easy to relate directly to behavioral reaction times because their analysis requires pooling together responses from a very large number of trials. The problem is that while a significant effect may be detected in the pooled data at a given latency, this does not mean that enough information could be extracted in real time on a single trial as would be required to initiate a behavioral response. However, a recent study of local field potential recordings obtained from occipito-temporal cortex in human epileptic patients reported that even in humans, information about object category can reliably be derived on a single trial basis from only 100 ms after stimulus onset (Liu, Agam, Madsen, & Kreiman, 2009). The study also showed that face-selective intracerebral responses were remarkably invariant to changes in size, position, and viewpoint, even at such short latencies. While these neurophysiological studies provide clear evidence that information about the presence of (for example) a face can *potentially* be extracted from brain activity shortly after stimulus onset, the current behavioral data goes further by demonstrating that such information can indeed be used by the brain to control behavior. Furthermore, while information can be extracted from intracortical potentials in humans from 100 ms (Liu et al., 2009), this does not imply that the information is necessarily already visible in neuronal firing. For example, in monkey IT, the earliest face-selective firing has been reported at latencies of 70–90 ms (Oram & Perrett, 1992), but deflections in intracerebrally recorded potentials are seen from as early as 50 ms (Schroeder, Mehta, & Givre, 1998). Thus, it might be that face-selective firing in the brain regions studied by Liu et al. might not occur until appreciably later, perhaps 120–30 ms, which would be well after the earliest face-selective saccades reported here.

While it may seem natural to assume that the processing required to initiate these fast responses involves the ventral cortical processing stream, it is important to realize that there is little reason to exclude subcortical processing pathways. For example, there is certainly evidence that face information can be processed in subcortical structures such as the amygdala, and there is evidence that visual information can reach the amygdala via the superior colliculus and the pulvinar (Johnson, 2005). Much of the evidence for subcortical processing has come from work with emotional faces and fear-inducing stimuli (Ohman, Carlsson, Lundqvist, & Ingvar, 2007), but it seems clear that the fast saccadic responses that we describe here are not restricted to faces with emotional expressions. Nevertheless, there is no strong argument for excluding the possibility that at least some of the information needed for face detection might have a subcortical origin.

One further result from the monkey neurophysiology that seems particularly important in the present context is the finding that the onset latencies of neurons in

inferotemporal cortex can vary significantly depending on the stimulus. Kiani, Esteky, and Tanaka (2005) reported that onset latencies of responses to primate and human faces are roughly 20 ms earlier than to faces of other animals. This difference was seen even though the total amount of activity is similar for the two types of stimulus, demonstrating that it is not simply that the neurons respond better to primate and human faces.

This result suggests an interesting possibility that could go some way toward explaining the remarkably rapid saccadic responses reported here. Suppose that when a face and a vehicle are simultaneously presented in the left and right visual fields, the neurons in the ventral stream contralateral to the face fire 20 ms earlier than the neurons responding to the vehicle. This may produce an imbalance in the levels of activation, which could result in differential activation in areas involved in saccade initiation such as the Frontal Eye Field (FEF) and the Lateral Intraparietal Area (LIP). It is as if the face would have a higher salience than other stimuli, simply because of this difference in onset latency.

If there is indeed a difference in the onset latency of neuronal responses to different types of stimuli, with the shortest latencies being seen for faces, then this might well explain why such short latency behavioral responses can be seen to faces. However, such an explanation would lead to the following hypothesis. Suppose that it is not possible to alter the latency of the neural responses under top-down control. In this case, we might expect that even if the subject was trying to direct their eyes toward the other stimulus, the latency advantage for faces would still result in a bias toward faces, at least for the fastest responses. This is precisely what we observed in Experiment 2, where we noted that when the subjects were instructed to saccade toward the vehicle, they nevertheless tend to saccade toward the face, at least when the saccades were initiated in under 140 ms.

A difference between manual and saccadic tasks?

The strong bias toward responding to faces reported here is not something that we have seen in previous studies using manual responses. These earlier studies using similar visual stimuli reported that subjects could easily shift between one target category and another from block to block. This was seen for both animals and means of transport (VanRullen & Thorpe, 2001a) and for animal and human faces (Rousselet et al., 2003). However, in these studies using a manual go/no-go task, even the very fastest responses are never seen earlier than about 250–300 ms following stimulus onset, considerably longer than the saccadic responses reported here. This raises the possibility that the extra processing time available in the manual task allows the subjects to generate a behavioral response that is fully under top-down control. In contrast, eye movements may be generated without allowing enough time for complete modulation of the behavior. This is clear from the results of Experiment 2 in which participants tried to selectively make saccades toward the vehicle. This pattern of results is precisely what would be expected if the visual system had an in-built bias in favor of faces, which would be evident even under conditions where the task requires the participants to saccade elsewhere.

The latency at which the saccades start to be modulated by the task requirements could be related to neurophysiological studies of attentional effects on neuronal responses. In studies of visual responses, it has repeatedly been noted that the initial transient part of the response tends to be relatively fixed and that task-related modulations only start to be clear after a further delay (Roelfsema et al., 2007; Treue, 2001). It is therefore possible that the difference in latency between the earliest saccades to faces (100–110 ms) and the earliest point at which subjects can start to reliably make saccades toward the vehicle target (around 140–150 ms) could reflect the duration of this initial period of the neural response, which appears to be relatively insensitive to top-down task modulation.

Once this top-down modulation has started, a decision mechanism that depended on the amount of cumulated activity originating from the two parts of the visual field would progressively become more and more strongly biased toward the intended target. As a consequence, saccades that are initiated at relatively long latencies could be reliably made in the direction of the target, even when the target is a vehicle. In the case of behavioral responses that have even longer latencies, such as a manual go/no-go response, the accumulation of activity will be sufficient to ensure that the response can be made to whatever target category is currently in use.

The fact that the effectiveness of top-down task-dependent modulation is not fixed, but varies depending on the latency at which the saccades are initiated, means that the saccadic choice task can be used to track the time course of top-down effects, something that may be difficult or even impossible using conventional manual reaction time methodologies. By the time a manual response is initiated, the brain has had plenty of time to complete a wide range of operations, including attentional modulation. In contrast, the very fastest saccadic responses appear to occur before these modulatory effects had time to go to completion. Until now, the only methods that have been able to investigate this early time window (100–150 ms) have been techniques such as EEG, MEG, and single-cell recording, but the behavioral significance of such effects have been difficult to establish.

What makes faces special?

It appears that faces may be in a class of their own in their ability to trigger very fast saccades, and a natural

question concerns the origin of this advantage. One possibility is that faces are special because we have a great deal of expertise in processing faces from an early age. Indeed, faces occupy a very important part of the visual environment of the newborn child (Sinha, Balas, & Ostrovsky, 2007), and this increased exposure could lead to the development of selective mechanisms using unsupervised learning (see, for example, Masquelier & Thorpe, 2007). On the other hand, there is considerable evidence that humans have an innately specified bias in favor of face stimuli (Johnson, Dziurawiec, Ellis, & Morton, 1991; Johnson & Mareschal, 2001), and recent work suggests that this preference carries across to photographs of chimpanzees with which we have had far less experience (Taubert, 2009). Another recent study showed that in a change-blindness paradigm, we are much more likely to notice a change involving an animal than one involving a vehicle, even when the stimuli have been matched for size within the image (New, Cosmides, & Tooby, 2007). The authors interpreted their results as favoring an ancestral priority for animals, including humans. Irrespective of the origin of this face bias, it will undoubtedly have a major impact on the way in which humans explore the visual environment.

Selectivity mechanisms

The final point concerns the mechanisms that can allow face-selective responses to be produced so rapidly. Our ability to initiate directed saccades toward faces as early as 100–110 ms after stimulus onset clearly leaves little time for anything other than a feed-forward pass. This point is strengthened by the fact that reaction times in the choice task are only marginally longer than in a simple detection task (see Experiment 3), meaning that the processing overhead associated with detecting the presence of a face in an image can be no more than a few tens of milliseconds. It has been known for some time that a single wave of spikes can be sufficient not only for face detection (VanRullen, Gautrais, Delorme, & Thorpe, 1998) but also for face identification (Delorme & Thorpe, 2001). There is further recent evidence that purely feed-forward hierarchical processing mechanisms may be sufficient to account for at least some forms of rapid categorization (Serre et al., 2007). In the field of computer vision, considerable progress has been made in developing algorithms for detecting and localizing faces in natural images (Hjelmas & Low, 2001; Viola & Jones, 2004), some of which have found their way into consumer products such as digital cameras.

Indeed, the shortest saccadic reaction times that we report here appear to be so fast that there may even not be enough time to complete the first feed-forward pass through the ventral processing stream. Remember that since we need to allow around 20 ms for response initiation, it seems inevitable that face-selective mechanisms must become activated from as early as 80 ms following stimulus onset. Even in monkeys, this would correspond to the earliest responses in inferotemporal neurons, but it needs to be remembered that monkeys are known to be substantially faster than humans on virtually all behavioral tasks, probably because their smaller heads lead to a reduction in conduction delays.

Given these constraints, it seems likely that the visual system will effectively try to make use of any available cues to the presence of a face in the image, especially those that can be extracted early on during visual processing. Some recent research suggests that this may indeed be the case. For example, Dakin and Watt (2009) have shown how the horizontal orientation structure in the human face provides a form of "bar code" that can be used for judgments of face identity. Further evidence for the use of very low level heuristics comes from another study from our laboratory, that also made use of the saccadic choice task, two images with varying contrast were presented on the left and right, and participants were required to saccade to the one with the highest contrast (Honey, Kirchner, & VanRullen, 2008). One of the images was a face, and the other one a vehicle—as in the experiments reported here. They found a very strong face bias in that even when the images had the same contrast, 70% of the saccades were made toward the faces. They then showed that at least some of this bias was still present even when the image was completely phase scrambled in the Fourier domain—that is, the images retained the same spatial frequency and orientation structure as the original images. This result supports the idea that rudimentary and quickly accessible information could explain a part of the bias we observed. Interestingly, a similar bias could also explain the tendency of faces to pop-out in multiple search arrays (Vanrullen, 2006). The underlying idea is that the visual system could use a simple heuristic characteristic of faces (for example, a pattern of high energy on both horizontal high and vertical low spatial frequencies), can be computed very rapidly, and used to generate useful selectivity early on in visual processing.

Conclusions

We have shown using a Saccadic Choice Task that humans can make very rapid saccades toward images containing faces. The earliest reliable saccades can be seen from as little as 100 to 110 ms after stimulus onset. While early face-selective electrical activity has been reported in a number of previous studies, this is the first time that it has been clear that such early selective responses could have a clear impact on behavior. The study also shows that the Saccadic Choice Task provides an experimental tool for studying very early processing in

the visual system, in a time window previously only accessible using electrophysiological techniques.

Acknowledgments

H. K. was supported by a European grant "Decisions In Motion" and by the ANR project "Hearing in Time". S. M. C. is supported by a grant from the Délégation Générale pour l'Armement. The research was also financed by the CNRS and by the ANR "Natstats" Project.

Commercial relationships: none.
Corresponding author: Simon Thorpe.
Email: simon.thorpe@cerco.ups-tlse.fr.
Address: Centre de Recherche Cerveau and Cognition, CNRS, University Toulouse 3, Toulouse, France.

References

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

Bindemann, M., Burton, A. M., Hooge, I. T., Jenkins, R., & de Haan, E. H. (2005). Faces retain attention. *Psychonomic Bulletin & Review, 12,* 1048–1053. [PubMed]

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Gaze selection in complex social scenes. *Visual Cognition, 16,* 341–355.

Braeutigam, S., Bailey, A. J., & Swithenby, S. J. (2001). Task-dependent early latency (30–60 ms) visual processing of human faces and other objects. *Neuroreport, 12,* 1531–1536. [PubMed]

Brown, V., Huey, D., & Findlay, J. M. (1997). Face detection in peripheral vision: Do faces pop out? *Perception, 26,* 1555–1570. [PubMed]

Buswell, G. T. (1935). *How people look at pictures: A study of the psychology of perception in art*. Chicago: University of Chicago Press.

Butler, S., Gilchrist, I. D., Burt, D. M., Perrett, D. I., Jones, E., & Harvey, M. (2005). Are the perceptual biases found in chimeric face processing reflected in eye-movement patterns? *Neuropsychologia, 43,* 52–59. [PubMed] [Article]

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in neural information processing systems* (vol. 20, pp. 241–248). Cambridge, MA: MIT Press.

Cohen, J. D., & Tong, F. (2001). Neuroscience. The face of controversy. *Science, 293,* 2405–2407. [PubMed]

Crouzet, S., Kirchner, H., & Thorpe, S. J. (2008). Saccading towards faces in 100 ms: What's the secret? *Perception, 37*(Supplement), 119–120.

Dakin, S. C., & Watt, R. J. (2009). Biological "bar codes" in human faces. *Journal of Vision, 9*(4):2, 1–10, http://journalofvision.org/9/4/2/, doi:10.1167/9.4.2. [PubMed] [Article]

Delorme, A., & Thorpe, S. J. (2001). Face identification using one spike per neuron: Resistance to image degradations. *Neural Networks, 14,* 795–803. [PubMed] [Article]

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science, 293,* 2470–2473. [PubMed]

Drewes, J., Trommershaeuser, J., & Gegenfurtner, K. R. (2009). The effect of context on rapid animal detection [Abstract]. *Journal of Vision, 9*(8):1177, 1177a, http://journalofvision.org/9/8/1177/, doi:10.1167/9.8.1177.

Eimer, M., & Holmes, A. (2002). An ERP study on the time course of emotional face processing. *Neuroreport, 13,* 427–431. [PubMed]

Einhauser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision, 8*(2):2, 1–19, http://journalofvision.org/8/2/2/, doi:10.1167/8.2.2. [PubMed] [Article]

Epshtein, B., Lifshitz, I., & Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 14298–14303. [PubMed] [Article]

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience, 13,* 171–180. [PubMed]

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review, 10,* 482–498. [PubMed]

Fischer, B., & Weber, H. (1993). Express saccades and visual attention. *Behavioral and Brain Sciences, 16,* 553–610.

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception, 37,* 571–583. [PubMed]

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal

lobe. *Nature Neuroscience, 12,* 1187–1196. [PubMed] [Article]

George, N., Jemel, B., Fiori, N., & Renault, B. (1997). Face and shape repetition effects in humans: A spatio-temporal ERP study. *Neuroreport, 8,* 1417–1423. [PubMed]

Gilchrist, I. D., & Proske, H. (2006). Anti-saccades away from faces: Evidence for an influence of high-level visual processes on saccade programming. *Experimental Brain Research, 173,* 708–712. [PubMed]

Guyonneau, R., Kirchner, H., & Thorpe, S. J. (2006). Animals roll around the clock: The rotation invariance of ultrarapid visual processing. *Journal of Vision, 6*(10):1, 1008–1017, http://journalofvision.org/6/10/1/, doi:10.1167/6.10.1. [PubMed] [Article]

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293,* 2425–2430. [PubMed]

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4,* 223–233. [PubMed]

Hemond, C. C., Kanwisher, N. G., & Op de Beeck, H. P. (2007). A preference for contralateral stimuli in human object- and face-selective cortex. *PLoS One, 2,* e574. [PubMed] [Article]

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7,* 498–504. [PubMed] [Article]

Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research, 45,* 1707–1724. [PubMed] [Article]

Hershler, O., & Hochstein, S. (2006). With a careful look: Still no low-level confound to face pop-out. *Vision Research, 46,* 3028–3035. [PubMed] [Article]

Hjelmas, E., & Low, B. K. (2001). Face detection: A survey. *Computer Vision and Image Understanding, 83,* 236–274.

Honey, C., Kirchner, H., & VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *Journal of Vision, 8*(12):9, 1–13, http://journalofvision.org/8/12/9/, doi:10.1167/8.12.9. [PubMed] [Article]

Hung, C. P., Kreiman, G., Poggio, T., & Dicarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310,* 863–866. [PubMed]

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences of the United States of America, 96,* 9379–9384. [PubMed] [Article]

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506. [PubMed] [Article]

Jacques, C., & Rossion, B. (2009). The initial representation of individual faces in the right occipito-temporal cortex is holistic: Electrophysiological evidence from the composite face illusion. *Journal of Vision, 9*(6):8, 1–16, http://journalofvision.org/9/6/8/, doi:10.1167/9.6.8. [PubMed] [Article]

Jeffreys, D. A. (1989). A face-responsive potential recorded from the human scalp. *Experimental Brain Research, 78,* 193–202. [PubMed]

Johnson, J. S., & Olshausen, B. A. (2003). Time course of neural signatures of object recognition. *Journal of Vision, 3*(7):4, 499–512, http://journalofvision.org/3/7/4/, doi:10.1167/3.7.4. [PubMed] [Article]

Johnson, J. S., & Olshausen, B. A. (2005). The earliest EEG signatures of object recognition in a cued-target task are postsensory. *Journal of Vision, 5*(4):2, 299–312, http://journalofvision.org/5/4/2/, doi:10.1167/5.4.2. [PubMed] [Article]

Johnson, M. H. (2005). Subcortical face processing. *Nature Reviews. Neuroscience, 6,* 787–798. [PubMed]

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40,* 1–19. [PubMed] [Article]

Johnson, M. H., & Mareschal, D. (2001). Cognitive and perceptual development during infancy. *Current Opinion in Neurobiology, 11,* 213–218. [PubMed] [Article]

Kano, F., & Tomonaga, M. (2009). How chimpanzees look at pictures: A comparative eye-tracking study. *Proceedings of the Royal Society B: Biological Sciences, 276,* 1949–1955. [PubMed]

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3,* 759–763. [PubMed]

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17,* 4302–4311. [PubMed] [Article]

Kiani, R., Esteky, H., & Tanaka, K. (2005). Differences in onset latency of macaque inferotemporal neural responses to primate and non-primate faces. *Journal of Neurophysiology, 94,* 1587–1596. [PubMed] [Article]

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research, 46,* 1762–1776. [PubMed] [Article]

Langton, S. R., Law, A. S., Burton, A. M., & Schweinberger, S. R. (2008). Attention capture by faces. *Cognition, 107,* 330–342. [PubMed] [Article]

Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron, 62,* 281–290. [PubMed]

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: An MEG study. *Nature Neuroscience, 5,* 910–916. [PubMed]

Masquelier, T., & Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Computational Biology, 3,* e31. [PubMed] [Article]

McCarthy, G., Puce, A., Belger, A., & Allison, T. (1999). Electrophysiological studies of human face perception. II: Response properties of face-specific potentials generated in occipitotemporal cortex. *Cerebral Cortex, 9,* 431–444. [PubMed]

Mormann, F., Kornblith, S., Quiroga, R. Q., Kraskov, A., Cerf, M., Fried, I., et al. (2008). Latency and selectivity of single neurons indicate hierarchical processing in the human medial temporal lobe. *Journal of Neuroscience, 28,* 8865–8872. [PubMed] [Article]

Mouchetant-Rostaing, Y., & Giard, M. H. (2003). Electrophysiological correlates of age and gender perception on human faces. *Journal of Cognitive Neuroscience, 15,* 900–910. [PubMed]

Mouchetant-Rostaing, Y., Giard, M. H., Bentin, S., Aguera, P. E., & Pernier, J. (2000). Neurophysiological correlates of face gender processing in humans. *European Journal of Neuroscience, 12,* 303–310. [PubMed]

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences, 104,* 16598–16603. [PubMed] [Article]

Ohman, A., Carlsson, K., Lundqvist, D., & Ingvar, M. (2007). On the unconscious subcortical origin of human fear. *Physiology Behavior, 92,* 180–185. [PubMed] [Article]

Oram, M. W., & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology, 68,* 70–84. [PubMed]

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42,* 107–123. [PubMed] [Article]

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45,* 2397–2416. [PubMed] [Article]

Puce, A., Allison, T., Gore, J. C., & McCarthy, G. (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology, 74,* 1192–1199. [PubMed]

Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure–ground mechanisms provide structure for selective attention. *Nature Neurosciences, 10,* 1492–1499. [PubMed] [Article]

Ro, T., Russell, C., & Lavie, N. (2001). Changing faces: A detection advantage in the flicker paradigm. *Psychological Science, 12,* 94–99. [PubMed]

Roelfsema, P. R., Tolboom, M., & Khayat, P. S. (2007). Different processing phases for features, figures, and selective attention in the primary visual cortex. *Neuron, 56,* 785–792. [PubMed]

Rossion, B., & Jacques, C. (2008). Does physical interstimulus variance account for early electrophysiological face sensitive responses in the human brain? Ten lessons on the N170. *Neuroimage, 39,* 1959–1979. [PubMed] [Article]

Rousselet, G. A., Mace, M. J., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of Vision, 3*(6):5, 440–455, http://journalofvision.org/3/6/5/, doi:10.1167/3.6.5. [PubMed] [Article]

Rousselet, G. A., Mace, M. J., & Fabre-Thorpe, M. (2004). Animal and human faces in natural scenes: How specific to human faces is the N170 ERP component? *Journal of Vision, 4*(1):2, 13–21, http://journalofvision.org/4/1/2/, doi:10.1167/4.1.2. [PubMed] [Article]

Rousselet, G. A., Mace, M. J., Thorpe, S. J., & Fabre-Thorpe, M. (2007). Limits of event-related potential differences in tracking object processing speed. *Journal of Cognitive Neuroscience, 19,* 1241–1258. [PubMed]

Schroeder, C. E., Mehta, A. D., & Givre, S. J. (1998). A spatiotemporal profile of visual system activation revealed by current source density analysis in the awake macaque. *Cerebral Cortex, 8,* 575–592. [PubMed]

Sergent, J., Ohta, S., & MacDonald, B. (1992). Functional neuroanatomy of face and object processing. A positron emission tomography study. *Brain: A Journal of Neurology, 115,* 15–36. [PubMed]

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 6424–6429. [PubMed] [Article]

Sinha, P., Balas, B., & Ostrovsky, Y. (2007). Discovering faces in infancy [Abstract]. *Journal of Vision, 7*(9):569, 569a, http://journalofvision.org/7/9/569/, doi:10.1167/7.9.569.

Tarr, M. J., & Gauthier, I. (2000). FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience, 3,* 764–769. [PubMed]

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45,* 643–659. [PubMed] [Article]

Taubert, J. (2009). Chimpanzee faces are "special" to humans. *Perception, 38,* 343–356. [PubMed]

Theeuwes, J., & Van der Stigchel, S. (2006). Faces capture attention: Evidence from inhibition of return. *Visual Cognition, 13,* 657–665.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381,* 520–522. [PubMed]

Thorpe, S., & Imbert, M. (1989). Biological constraints on connectionist modelling. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, & L. Steels (Eds.), *Connectionism in perspective* (pp. 63–93). Amsterdam, The Netherlands: Elsevier.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends in Neurosciences, 24,* 295–300. [PubMed] [Article]

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience, 31,* 411–437. [PubMed] [Article]

Vanrullen, R. (2006). On second glance: Still no high-level pop-out effect for faces. *Vision Research, 46,* 3017–3027. [PubMed] [Article]

VanRullen, R., Gautrais, J., Delorme, A., & Thorpe, S. (1998). Face processing using one spike per neurone. *BioSystems, 48,* 229–239. [PubMed]

VanRullen, R., & Thorpe, S. J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception, 30,* 655–668. [PubMed]

VanRullen, R., & Thorpe, S. J. (2001b). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience, 13,* 454–461. [PubMed]

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42,* 2593–2615. [PubMed] [Article]

Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision, 57,* 137–154.

Vuilleumier, P. (2000). Faces call for attention: Evidence from patients with visual extinction. *Neuropsychologia, 38,* 693–700. [PubMed] [Article]

Yarbus, A. F. (1967). *Eye movements and vision*. New York: Plenum Press.

2.2.3 Résumé des principaux résultats

Cette série d'études a donc permis de montrer, en utilisant la tâche de choix saccadique initiée par Kirchner et Thorpe (2006), que des sujets humains pouvaient produire des réponses comportementales sélectives très rapides basées sur la reconnaissance de différentes catégories d'objets. Nous avons surtout pu mettre en évidence des différences de temps de traitement entre les catégories d'objets qui étaient invisibles en utilisant un protocole de réponse manuelle. Le résultat le plus important de cette étude reste cependant la capacité pour les sujets humains d'initier des saccades sélectives vers les visages avec un temps de réaction minimum de seulement 100 ms. Des temps de réaction si courts suggèrent que le système visuel n'a certainement pas le temps d'effectuer un traitement complet depuis la rétine jusqu'à IT. Différentes alternatives sont possibles pour expliquer des temps de réaction si courts.

2.2.4 Différentes explications possibles

Compte tenu des temps de réaction très courts observés dans les expériences citées précédemment, il est difficile d'imaginer que l'information visuelle puisse emprunter le trajet classique associé à la reconnaissance d'objet (rétine → LGN → V1 → V2 → V4 → IT). Deux alternatives peuvent être envisagées : l'information visuelle peut soit emprunter un raccourci, soit ne pas avoir besoin d'aller jusqu'au bout du chemin. Pour cette seconde possibilité, le système visuel utiliserait alors un état précoce de l'information pour réaliser la tâche. Ces deux hypothèses ne sont pas antinomiques et peuvent tout à fait être combinées. Je passerai donc en revue les différentes possibilités, notamment sous-corticales, pouvant être le support d'un transport si rapide de l'information. Ceci nous amènera à étudier dans la section 2.3 une hypothèse précise concernant l'état précoce de l'information que pourrait utiliser le système pour gagner du temps.

Le cerveau est un organe infiniment complexe, et de nombreuses connexions existent entre la plupart des aires. Il est donc impossible de faire une revue exhaustive de tous les chemins que peut prendre l'information visuelle pour supporter la détection ultra-rapide d'objets dans le champ visuel. J'évoquerai donc ici les trois possibilités les plus plausibles : un chemin purement sous-cortical, un passage par la voie dorsale, et un passage accéléré dans la voie ventrale.

Un chemin purement sous-cortical (et archaïque ?)

Une première possibilité repose sur un nombre conséquent d'études suggérant l'existence d'un chemin dédié aux visages purement sous-cortical (Johnson, 2005). Ce chemin impliquerait principalement les colliculi supérieurs, le pulvinar et l'amygdale à partir de l'information de la voie magnocellulaire. Cette activité purement sous-corticale expliquerait qu'une activité sélective aux visages puisse être détectée en MEG avant

même les premières activations de V1 (Bailey *et al.*, 2005; Braeutigam *et al.*, 2001). Les caractéristiques de cette route seraient qu'elle est rapide, opère à partir des basses fréquences spatiales, et module les traitements corticaux (Johnson, 2005). Elle pourrait expliquer la préférence des nouveaux-nés pour les formes ressemblant à des visages. Johnson propose également que de par son probable effet de modulation de l'activité corticale (Keightley *et al.*, 2003), elle pourrait être à la base de l'émergence de la zone corticale dédiée aux visages qu'est la FFA (mais ce dernier point reste très hypothétique, Johnson, 2005). Chez l'adulte, ce chemin pourrait donc être une voie "*quick & dirty*" permettant la détection rapide (dans l'ensemble du champ visuel) de stimuli importants. Cette reconnaissance spécifique des visages chez l'humain pourrait trouver un écho chez d'autres espèces pour la reconnaissance de leurs congénères. L'espèce des grenouilles semble ainsi avoir développée des réseaux sous-corticaux innés dans ce but (Sewards et Sewards, 2002). Le système dont nous parlons ici pourrait donc en être un dérivé.

Connexions colliculi → voie dorsale

Une voie de traitement alternative est constituée par des connexions qui vont directement de la rétine vers les colliculi supérieurs, puis font relais dans le pulvinar et le LGN avant d'atteindre la voie dorsale dans le cortex, plus précisément vers V3 et MT (Lyon *et al.*, 2010). Il ne semble pas y avoir de connexion directe vers la voie ventrale (pas de connexions vers V2 ou V4). La majorité de ces connexions allant vers MT et les neurones impliqués au niveau des colliculi étant semblables à des neurones magnocellulaires, il semble que cette voie pourrait être très majoritairement impliquée dans la détection extrêmement rapide de mouvements (et pourrait donc être un support pour les saccades rapides vers les mouvements d'expansion que nous avons observées dans une de nos études [1]). Elle pourrait ainsi être à l'origine des capacités visuelles résiduelles dans les cas de vision aveugle (Sanders *et al.*, 1974). Cependant, l'implication de cette voie pour le traitement de la forme et des objets n'a pas été démontré pour le moment. Une alternative beaucoup plus plausible réside dans la présence de neurones sélectifs à la forme dans les aires visuo-motrices comme FEF et LIP (cf section 1.3.5) qui pourraient être activés très vite après un passage par le cortex primaire sans passer par la voie ventrale proprement dite.

Un parcours accéléré de la voie ventrale vers FEF

Une troisième alternative est un passage accéléré dans la voie ventrale. Plusieurs possibilités sont envisageables pour raccourcir le trajet. Selon le modèle hiérarchique

1. S J Thorpe, H Kirchner, S Crouzet, P Bayerl, H Neumann (2008). Processing times for optic flow patterns measured by the saccadic choice task. Perception 37 ECVP Abstract Supplement, page 40.

classique, on aurait 6 étapes nécessaires : rétine → LGN → V1 → V2 → V4 → FEF. Il est possible que des connexions directes de V1 à V4 (Felleman et Van Essen, 1991) ou (encore plus rapide) du LGN à V4 (Fries, 1981) permettent de raccourcir très fortement ce chemin. Même si ces différentes connexions restent largement minoritaires, elles existent tout de même et pourraient jouer un rôle pour les réponses très rapides.

Ces différentes alternatives ne sont pas exclusives, bien au contraire. Comme le souligne Jean Bullier, différentes informations sont traitées à différentes vitesses et par différentes voies, et le cerveau les exploite certainement toutes au maximum dans le but de réaliser une tâche donnée (Bullier, 2004). On aurait donc un multitude de systèmes visuels qui agissent en parallèle, tout en inter-agissant largement les uns avec les autres.

2.3 Quel rôle pour les indices bas-niveau dans les traitements ultra-rapides ?

Au lieu d'utiliser un raccourci, le système visuel pourrait utiliser ce qui pourrait être assimilé à un *proxy* (un intermédiaire) pour réaliser la tâche. C'est à dire se baser sur une information disponible rapidement et lui permettant de remplir l'objectif fixé. Selon l'organisation hiérarchique, l'aire V4 jour un rôle d'intermédiaire entre le cortex primaire et les aires sélectives aux objets. L'information contenue dans l'activité des neurones de cette aire pourrait donc être accessible plus rapidement que celle des aires supérieures comme IT. De plus, il semble que cette aire pourrait jouer un rôle important dans l'exploration oculaire de scènes naturelles (Mazer et Gallant, 2003). Elle fait donc figure de candidate crédible pour être à la base des réponses rapides que l'on enregistre dans nos études.

Comme nous l'avons vu dans la section 1.3.1, la sélectivité des neurones de V4 est complexe, et semblent être décrite au mieux dans le domaine spectral (David *et al.*, 2006). Selon cette étude, les neurones de V4 auraient le plus souvent une sélectivité bi-modale à deux orientations associées chacune à une fréquence spatiale. la sélectivité à la fréquence spatiale serait aussi plus large que par exemple dans V1. Théoriquement, on pourrait donc trouver des neurones dans V4 (ou des groupes de neurones) sensibles à des combinaisons simples d'orientations spécifiques aux objets que l'on recherche. Ces détecteurs ne seraient pas aussi génériques que ceux de IT par exemple, mais pourraient constituer une base pour les réponses très rapides. Le système saccadique pourrait-il alors se baser sur ces détecteurs pour guider les saccades vers les visages ? Du fait du type de sélectivité des neurones de V4, une information qui pourrait ainsi jouer un rôle de *proxy* serait le spectre d'amplitude des images dans le domaine fréquentiel de Fourier. Cependant, cette hypothèse n'est pas strictement reliée à la sélectivité des

neurones de V4. Du fait de l'organisation générale du cortex visuel précoce (V1, V2, V4), le spectre d'amplitude d'une image présentée est représenté directement dans le pattern d'activation de l'ensemble de ces aires, ce type d'information est donc largement disponible si l'on prend l'information globale contenu dans ces aires.

Comme nous l'avons décrit dans la section 1.1.2, la description des images en terme de fréquences spatiales s'est avérée fructueuse pour la compréhension des mécanismes du système visuel. L'assimilation des premières étapes du système visuel à un traitement de Fourier remonte à plusieurs décennies (Campbell et Robson, 1968; Marr, 1982; Westheimer, 2001). On associe généralement l'amplitude aux caractéristiques bas-niveaux de l'image. Par opposition, le sens serait contenu dans le spectre de phase qui organise la positions des différentes sinusoïdes entre elles (Oppenheim et Lim, 1981). Cependant, Oliva et Torralba ont montré, il y a quelques années, que l'amplitude était aussi porteuse d'informations concernant la catégorie de la scène (Oliva et Torralba, 2001; Torralba et Oliva, 2003). Si elle contient effectivement de l'information, alors pourquoi le système visuel ne l'utiliserait-il pas ? Ceci a été montré de façon indirecte. Par exemple, une amorce basée sur le spectre d'amplitude peut biaiser la réponse des sujets dans une tâche de catégorisation (Guyader, 2004; Kaping *et al.*, 2007). Cependant, la grande majorité des études ayant testé son utilisation directe ont conclu que l'amplitude ne semblait pas être effectivement utilisée (voir Gaspar et Rousselet, 2009; Wichmann *et al.*, 2006, pour des exemples testant la catégorisation animal/non-animal), ceci même en utilisant une réponse saccadique (Wichmann *et al.*, 2010).

La détection de visages pourrait cependant être un cas à part. Il se pourrait par exemple que l'information d'amplitude spécifique des visages explique à elle seule les effets de *pop-out* des visages en recherche visuelle (VanRullen, 2006). Il a été montré récemment, dans une tâche de choix saccadique, que même lorsque la phase des images était brouillée au maximum (ne transportant ainsi plus aucune information sémantique à première vue), les saccades les plus rapides continuaient à être préférentiellement orientées vers l'image qui contenait le spectre d'amplitude du visage (Honey *et al.*, 2008). Ainsi, l'information d'amplitude des visages pourrait être très spécifique (Dakin et Watt, 2009; Keil, 2008; Nestor *et al.*, 2008). Le système visuel pourrait donc exploiter cette particularité pour réaliser la détection ultra-rapide mise en avant jusqu'ici.

2.3.1 Résumé de l'étude

Afin d'explorer l'utilisation des informations spectrales d'amplitude dans la détection rapide d'objet, nous avons comparé dans une première expérience les performances des sujets dans une tâche visage vs. véhicule (similaire à l'Expérience 2 de l'article 1), à leurs performances dans la même tâche, mais où l'on avait normalisé (moyenné) le spectre d'amplitude entre toutes les images. Dans cette condition normalisée, ils ne

pouvaient se baser que sur l'information de phase. Cette opération diminuait significativement leurs performances. Cependant, il a été montré récemment qu'une baisse de performance après ce type de normalisation n'attestait pas forcément de l'utilisation du spectre d'amplitude en soi, mais pouvait provenir du bruit ajouté globalement à l'image (Gaspar et Rousselet, 2009). Afin de tester plus précisément l'utilisation par les sujets de ces informations, nous nous sommes donc inspirés de l'étude de Gaspar & Rousselet, pour réaliser une deuxième expérience. Dans celle-ci, nous avons comparé les performances des sujets dans trois conditions différentes :

- ORI : images originales
- INV : les visages ont des amplitudes de véhicules et inversement
- SWA : les visages ont les spectres d'un autre visage, de même pour les véhicules

Dans la condition SWA, les images ont donc des spectres d'amplitude et de phase qui correspondent à la même catégorie, mais pas à la même image. L'apparence de l'image étant extrêmement sensible à la cohérence entre amplitude et phase, ces images paraissent très bruitées (au moins autant que celles de la condition INV en tout cas, ce qui est fondamental ici).

Lorsque la cible est le visage, les sujets ont des performances similaires dans les conditions ORI et SWA, et des performances inférieures dans la condition INV. Ce patron de résultat démontre l'utilisation de l'amplitude par les sujets pour détecter les visages, même si la phase reste fondamentale, comme en atteste le niveau de performance tout à fait respectable dans la condition INV.

2.3.2 Article 2 : Swap the face ! Use of amplitude spectrum to drive fast saccades

Swap the face!

Use of amplitude spectrum by the visual system to drive fast saccades

(en préparation)

Sébastien M. Crouzet[1,2] and Simon J. Thorpe[1,2]

[1] Université de Toulouse, UPS, Centre de Recherche Cerveau et Cognition, France

[2] CNRS, CerCo ; Toulouse, France

Corresponding Author:
Simon J. Thorpe
Centre de Recherche Cerveau et Cognition
Faculté de Médecine Rangueil
31062 Toulouse Cedex (FRANCE)

Tel: +5 62 17 28 03
Fax: +5 62 17 28 09
email: simon.thorpe@cerco.ups-tlse.fr

Abstract

When images of a face and a vehicle are flashed left and right of fixation, subjects can selectively saccade toward the face only 100 ms after image onset (Crouzet, Kirchner, & Thorpe, 2010). These fast responses probably do not allow enough time for a complete analysis of the image by the ventral stream. What sorts of information could be used for triggering such fast saccades? One possibility is that this ultra-rapid processing relies on relatively low-level amplitude spectrum (AS) information in the Fourier domain (Honey, Kirchner, & VanRullen, 2008). Thus, Experiment 1 showed that AS normalization in the task can significantly alter face detection performance. However, a decrease of performance following AS normalization does not prove that amplitude spectrum based information is used (Gaspar & Rousselet, 2009). In Experiment 2, following the Gaspar and Rousselet paper, we used a swapping procedure to clarify the role of AS information in fast object detection. Our experiment used 3 conditions: (i) original images, (ii) inverted, in which the face image has the AS of a vehicle, and the vehicle has the AS of a face, and (iii) swapped, where the face has the AS of another face, and the vehicle has the AS of another vehicle. The results showed very similar levels of performance in the original and swapped conditions, and a clear drop in the inverted condition. Thus, in the early temporal window offered by the saccadic choice task, the visual saccadic system does effectively make use of low-level AS information for fast face detection, even if phase information seems to still be more important.

Keywords: natural scenes, fast saccades, Fourier transform, amplitude spectrum, face detection

INTRODUCTION

The time needed to detect the presence of an object in a complex natural scene can be remarkably short (Kirchner & Thorpe, 2006; Potter, 1975; Thorpe, Fize, & Marlot, 1996). The Kirchner & Thorpe study showed that if you present two images left and right of a fixation to human subjects, they can selectively saccade on the image containing an animal as early as 120-130 ms after stimulus onset. Recently, it has been show that this processing time could be even shorter if targets are human faces, with reliable saccadic responses starting after only 100 ms when vehicles are used as distractors (Crouzet, et al., 2010). Additionally, this extremely fast processing is associated with a very strong bias toward faces, such that when subjects attempt to saccade toward another category of stimulus such as vehicles when faces are distractors, performance is particularly poor, especially when the saccades are initiated at short latencies.

This kind of extremely rapid processing puts very severe constraints on underlying visual processing. Given that the earliest saccades toward faces can be initiated at 100 ms, it follows that the brain mechanisms that trigger the response must be even earlier. It is often assumed that the delays in the oculomotor circuit leading to activation of the eye muscles are of the order of 20 ms or so, implying that the “decision” to move is presumably made at a latency of only 80 ms. Even a pure feedforward processing sweep from the retina to the human homologue of inferotemporal cortex (IT), where object selective neurons are found (Tanaka, 1996), might be too long given the latencies of single-cell recordings in monkeys (see Lamme & Roelfsema, 2000 for a review) and Local Field Potential (LFP) studies in humans (Liu, Agam, Madsen, & Kreiman, 2009). Instead, the visual system might base these rapid behavioral decisions on information that is only partially processed, leading to what might be termed "quick and dirty" processing.

A good candidate to be the basis of this "quick and dirty" processing could be amplitude spectrum information in the Fourier domain. The analogy between the early stages of processing in the visual system and Fourier analysis is long standing (Marr, 1982; see Westheimer, 2001 for an historical review). Early visual processing has often been described as a form of filtering operation (Campbell & Robson, 1968; Field, 1999; Marr, 1982). Although image semantics does not depend on spatial-frequency amplitude but rather on phase information (Oppenheim & Lim, 1981; Piotrowski & Campbell, 1982), the

Fourier spectral signatures of scenes have been used by computational models to infer scene categories (Oliva & Torralba, 2001; Torralba & Oliva, 2003). Indeed, it has been suggested that the human visual system can take advantage of these low-level natural image statistics to perform more complex tasks. For example, rapid image recognition can be biased by priming using information concerning the amplitude spectrum (for global scene properties: Guyader, Chauvin, Peyrin, Herault, & Marendaz, 2004; or animal detection: Kaping, Tzvetanov, & Treue, 2007). However, studies that have directly manipulated the target images for recognition suggest that it is not used for global scene categorization (Loschky & Larson, 2008; Loschky, et al., 2007) but see (Joubert, Rousselet, Fabre-Thorpe, & Fize, 2009) and animal detection (Gaspar & Rousselet, 2009; Wichmann, Braun, & Gegenfurtner, 2006), even when tested using fast saccadic responses (Felix A. Wichmann, Jan Drewes, Pedro Rosas, & K. R. Gegenfurtner, 2010).

However, face detection could be a special case. Amplitude information has been claimed to be responsible for face pop-out in visual search (Vanrullen, 2006), but see also (Hershler & Hochstein, 2006). Recently, a study explicitly tested the role of amplitude information for fast saccades toward faces. Using a similar design to Wichmann et al., but replacing the manual response by a saccadic choice task, they showed that amplitude information alone can bias fast saccadic responses toward faces (Honey, et al., 2008). Furthermore, numerous studies have already showed that faces have specific characteristics in the frequency domain (Dakin & Watt, 2009; Keil, 2008, 2009; Keil, Lapedriza, Masip, & Vitria, 2008; Nestor, Vettel, & Tarr, 2008). As a demonstration, a classifier based purely on amplitude information has a remarkably good level of performance when separating the set of images used in Crouzet et al., 2010 (85% on faces and 84% on vehicles, see Figure 1). This particularity could thus in principle be used by the visual saccadic system.

In the present study, using the same set of images than in an earlier study, we investigated the role of amplitude spectrum information in ultra-rapid detection of faces. Experiment 1 demonstrated a clear performance decrease in the absence of amplitude spectrum information. However, it is possible that this decrease would only be caused by the edge noise added to images and not the intended normalization of amplitude spectrum (Gaspar & Rousselet, 2009). Experiment 2, designed to address this issue, showed that edge noise cannot explained alone this result, but rather suggested that amplitude

information is effectively used to guide fast saccades, even if phase information is still the most important.

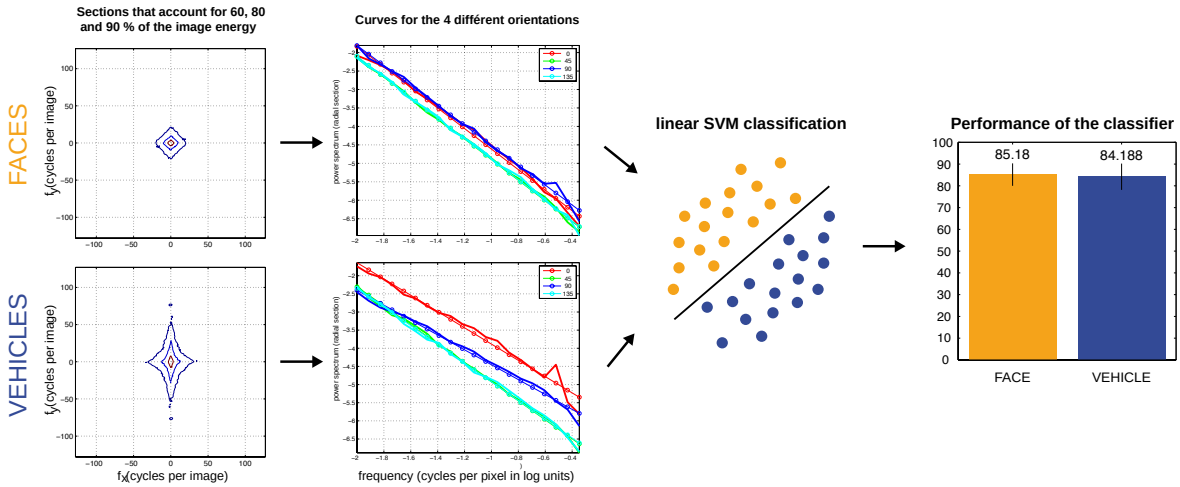

Figure 1 : Performance of a linear classifier on the face and vehicle images used in Crouzet et al., 2010. *After having resized images to 256x256, they were passed trough a hamming window function to remove boundary artifacts. The amplitude of the Fourier transform was then computed on each image. The resulting distribution of frequencies were divided into 4 bins of orientation (horizontal, vertical, and the two obliques), each one covering 45°. The distribution of frequencies for each orientation was then encoded by 20 points (from low to high frequencies, regularly spaced in log coordinates), resulting in 80 values representing the global features of each image used to feed the classifier. A linear SVM was then trained on half of the images (50% faces, 50% vehicles) and then tested on the other half. After 1000 cross-validation with random train and test subsets, we computed the mean performance of the classifier to correctly classify an image in its class. The error bars correspond to the standard deviation.The computing of the global and unlocalized features was based on the MATLAB script AverageAndPowerSpectrum.m (Torralba and Oliva, 2003), and the classification was done using the LIBSVM 2.9-1 for MATLAB (http://www.csie.ntu.edu.tw/~cjlin/libsvm/).*

GENERAL METHODS

Stimuli

200 photographs selected from the Corel Photolibrary database and downloaded from the Internet were used to set up two object categories of 100 natural scenes: human faces and vehicles. The same set has already been used in a recent study by our group (Crouzet, et al., 2010). All the images were converted to gray-scale and resized to 330*330 pixels. The global luminance (0.5) and contrast (RMS = 0.26) were set to be equal between images. All the image modifications were done using MATLAB.

Apparatus

Participants viewed the stimuli in a dimly lit room with their heads on a chin rest to maintain the viewing distance at 60 cm. Stimuli were presented on a IIYAMA Vision Master PRO 454 monitor with the screen resolution set to 800*600 pixels and a refresh rate of 100 Hz. The centers of the two images were always 8.6° from the fixation cross, resulting in a retinal size for each image of 14° by 14°. Stimuli presentation was done using Matlab and the Psychophysics Toolbox 3 (Brainard, 1997; D. Pelli, 1997). The background color is the only apparatus difference between the two experiments. In Experiment 1, a black background was used. In Experiment 2, it has been changed to a mid gray background, which seems to slightly improve subjects performance.

The saccadic choice task

A trial takes place as follows : Observers had to keep their eyes on a fixation cross which disappeared after a pseudo-random time interval (800-1600 ms). After a 200 ms-time gap, two natural scenes (one face and one vehicle) appeared on each side of the screen for 400 ms (see Figure 2). The task was to make a saccade as quickly and as accurately as possible to the side of the target.

Eye movement recording

Eye movements were monitored with an IView Hi-Speed eyetracker (SensoMotoric Instruments, Berlin, Germany). This infrared tracking system samples eye position at 240 Hz. Saccade detection was performed off-line using the saccade based algorithm of the SMI BeGaze Event Detection (Smeets & Hooge, 2003). Only the first saccade to enter one

of the 2 images after display onset was analyzed. Before each run, a 13-point calibration was performed.

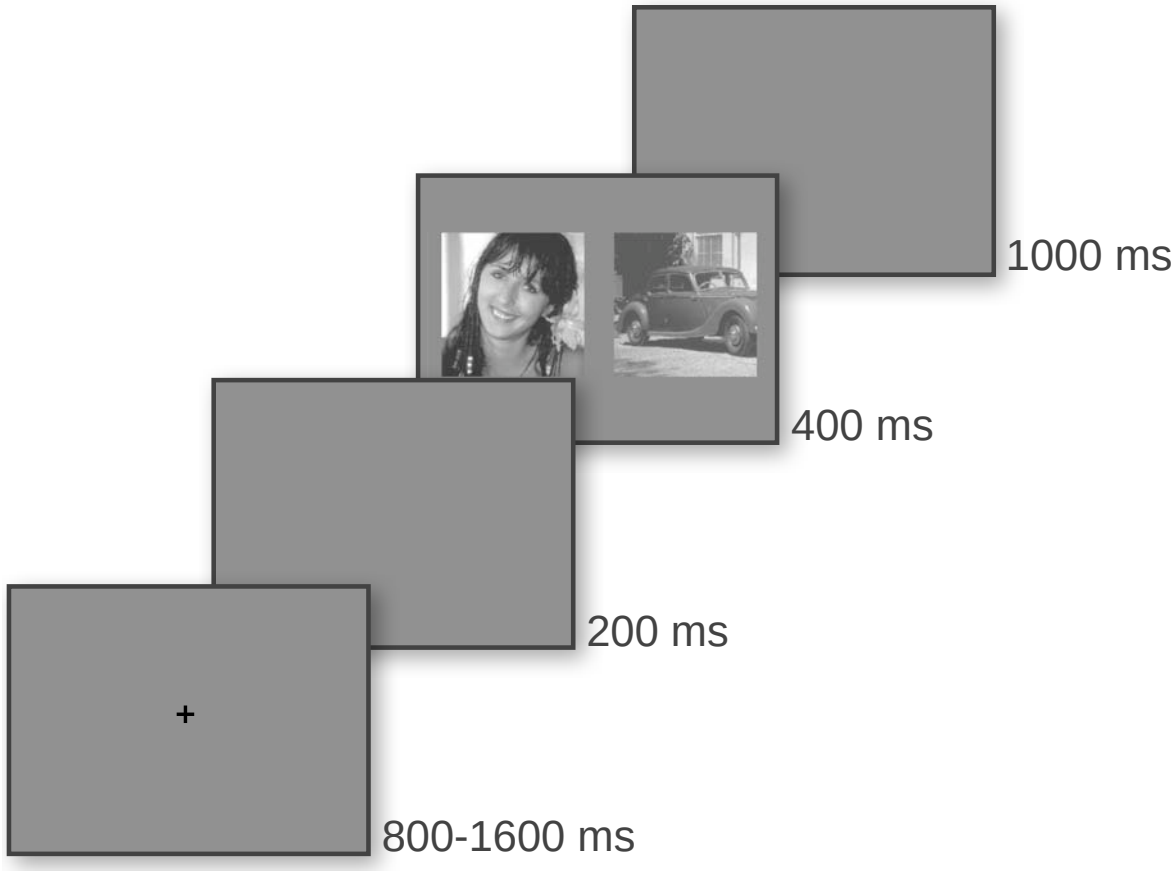

Figure 2: Protocol*: The saccadic choice task. A fixation cross is displayed for a pseudo-random time (between 800 and 1600 ms), then, after a 200 ms gap, 2 images (a target and a distractor) are displayed on the left and right. After a 1000 ms period, a new trial can start.*

EXPERIMENT 1

In order to test the influence of amplitude spectrum for the ultra-rapid detection of faces, we made this cue non informative by averaging it between images. This correspond to the normalized condition. In this case, the only information still available to discriminate between faces and vehicles is thus phase information. The performance of participants in this normalized condition was compared to their performance with original images.

Methods

Participants. 8 participants (6 males, mean age = 29.5, 2 left-handed) including the two authors gave written informed consent before participating in the experiment.

Amplitude spectrum normalization. In order to perform the normalization, we computed the mean amplitude spectrum over all images of the two categories. This mean amplitude spectrum was then combined with the original phase of each image, resulting in a second equivalent set of images which differed only by their phase information (Figure 3). This method averages the spatial frequency contents of all stimuli at each scale and orientation.

Design. Using a within-subject design, two tasks (saccading toward faces or toward vehicles) and two types of images (original, normalized) were tested here. The whole experiment was divided for each subject in two blocks (one for each task). In each block, runs alternated between the two types of images. For example, a subject would start with 8 runs of 50 trials where the target were faces — these runs alternating between original images and normalized images — followed by another 8 runs with vehicles as target, blocks and runs orders being counterbalanced between subjects. Indeed, each image was displayed two times in each condition (once on the left and once on the right hemifield), resulting in 200 trials per condition and per subject. Each block was preceded by a training session of 50 trials (25 with original images and then 25 with normalized images which were not used in the experiment).

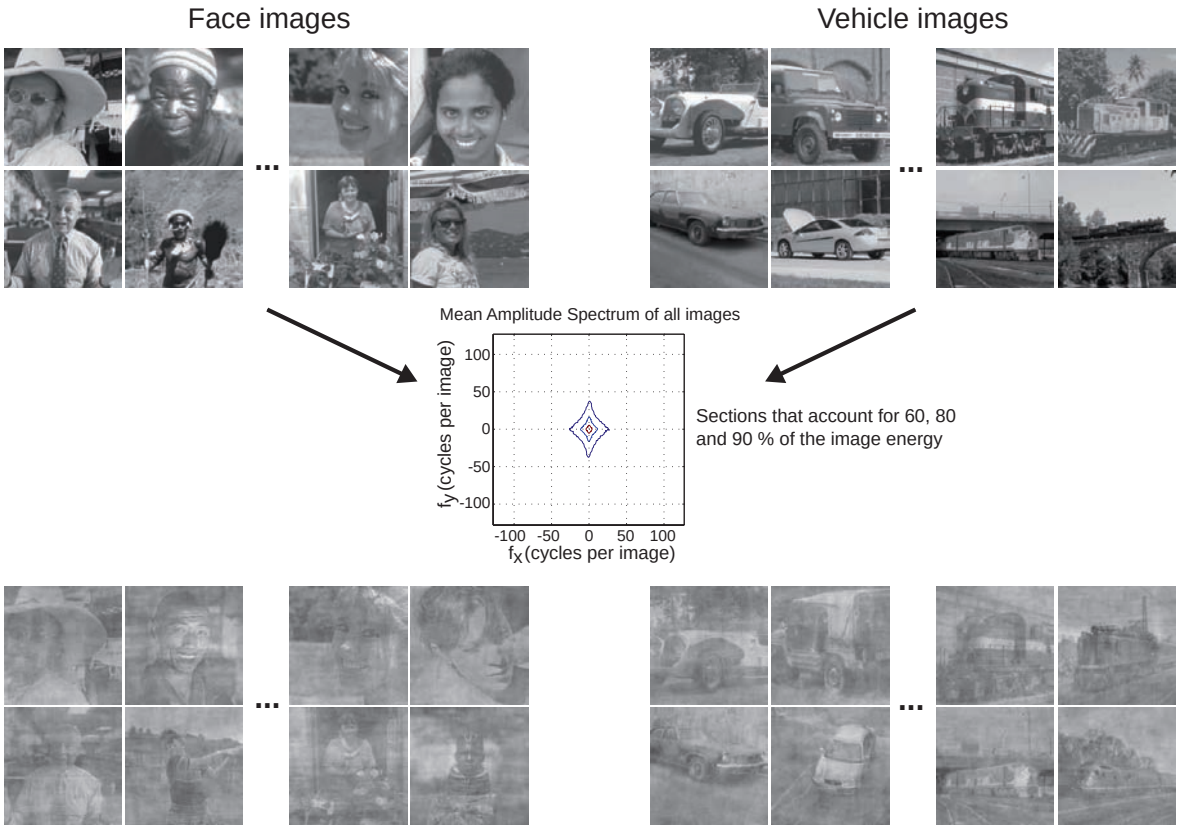

Figure 3: Stimuli of Experiment 1*. Examples of original images of faces and vehicles at the top. In the middle, the mean amplitude spectrum of faces and vehicles images is plotted using a representation where each section account for 60, 80 and 90 % of the image energy. This mean amplitude spectrum have then been applied to every image to create the normalized condition. Examples of images for this condition are at the bottom of the figure.*

Results

The results with original images (mean RT of 168 ms and 88.8% correct in the face task; 197 ms and 72.8% in the vehicle task) are very comparable to the values observed in a previous experiment by Crouzet et al., 2010 (Figure 4). In the normalized condition, when participants can only rely on phase information, they were still able to do the task very quickly and accurately (180 ms and 75% correct in the face task; 201 ms and 63.8% in the vehicle task). Using a 2-way ANOVA with factors Task (face, vehicle) and Type (original, normalized), the results showed that even though participants are able to do the task above chance in the normalized condition, their performance is globally lower (Type global effect: $F_{(1,7)}=55.65$, $p<.001$ for mean RT; $F_{(1,7)}=108.12$, $p<.001$ for accuracy). A result which is consistent with previous studies using this type of normalization (Gaspar &

Rousselet, 2009; Joubert, et al., 2009; Loschky & Larson, 2008; Loschky, et al., 2007; Wichmann, et al., 2006; Felix A. Wichmann, et al., 2010).

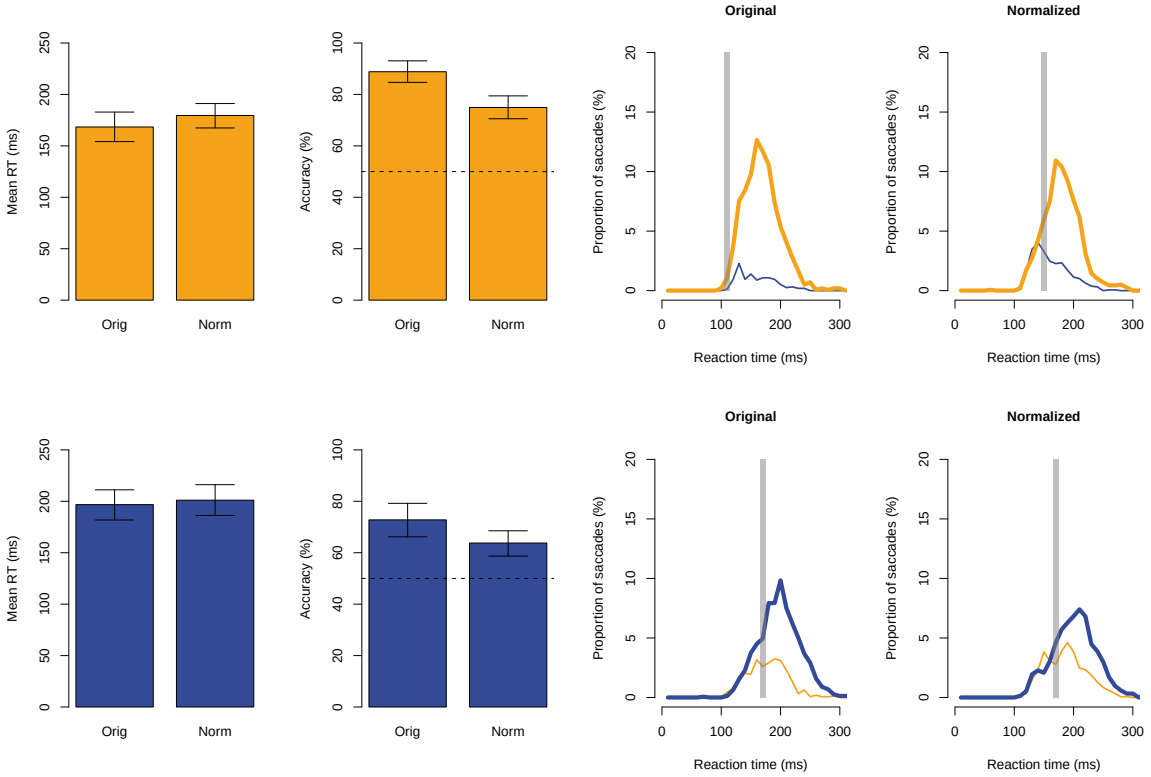

Figure 4: Results of Experiment 1. *On the top, results when participants had to saccade toward faces, on the bottom, results when participants had to saccade toward vehicles.* ***(a)*** *Mean RT and accuracy across subjects in the original (Orig) and normalized (Norm) conditions. Error bars represent bootstrapped 95% C.I. of the mean.* ***(b)*** *Distributions of RT in each condition and each task. Thick lines are for correct responses, thin lines for errors. Orange lines for saccades toward faces, blue lines for saccades toward vehicles. The transparent gray bar is placed according to the minimum RT (the first 10 ms bin of time where correct responses significantly outnumber incorrect ones using a* χ^2 *test).*

Furthermore, despite the normalization, the advantage for faces is still present although reduced in size (significant interaction on mean RT: $F_{(1,7)}=11.33$, $p<.01$), with an advantage for faces over vehicles of 21 ms and 11% accuracy in the normalized condition, 29 ms and 16% in the original one (Task global effect: $F_{(1,7)}=17.69$, $p<.01$ for mean RT; $F_{(1,7)}=27.65$, $p<.01$ for accuracy). A second conclusion is thus that the amplitude spectrum can explain a part of the bias toward faces but is not sufficient.

As a summary of Experiment 1, and as expected from previous studies, the amplitude spectrum normalization significantly decreases subjects performance. However, and as expected, this effect was larger in the face task than in the vehicle one. Even if phase information can be sufficient to induce a bias toward faces, amplitude information also plays a significant role in driving fast saccades.

EXPERIMENT 2

A recent study demonstrated that a performance decrease caused by amplitude spectrum normalization is not sufficient to claim that it is effectively used by the visual system to perform the task. Indeed, Gaspar and Rousselet showed that the performance decrease in an animal detection task caused by amplitude spectrum normalization could be explained by the edge noise added incidentally by the manipulation, rather than by the absence of amplitude information (Gaspar & Rousselet, 2009). To show that, they used an additional condition where animal images phase spectra were combined with animal images amplitudes, but always from different images (they did the same for non-animal images), resulting in images with phase and amplitude from the same category but from different individual images. The performance of subjects in this "swapped" condition was similar to the one in the "normalized" condition. This means that in a manual response animal detection task, what matters is not phase or amplitude information by themselves, but the interaction between the two.

However, it has been shown that the amplitude spectrum of faces alone could attract fast saccades in a saccadic choice task (Honey, et al., 2008) since even when the phase information was completely scrambled, there was still a bias with saccades toward faces. In the Experiment 2, we thus used two new image manipulations: (i) the inverted condition (INV), where the amplitude spectra of the face and vehicle images were exchanged and (ii) the swapped condition (SWA) where the amplitude spectra applied to each image were taken from another image of the same category. Thus, in all conditions, amplitude differences were still informative, but in the case of INV, they were inverted between categories. The only difference between SWA and ORI being that amplitude and phase of each image was not consistent. This results in 3 different conditions: original

(ORI), inverted (INV) and swapped (SWA) that will be used to investigate further the role of amplitude spectrum in guiding fast saccades toward faces.

The logic here is as follows. If the amplitude is not used by subjects to perform the task, the performance in the INV and SWA conditions should be equal, and certainly lower than in ORI because of higher edge noise (Gaspar & Rousselet, 2009). If the amplitude spectrum can be used independently of phase information, performance in the ORI and SWA conditions should be equal, INV being significantly lower. Our results clearly support the independence alternative for fast saccades.

Methods

Participants. 12 participants (7 males, mean age = 25.9, 1 left-handed) including the first author gave written informed consent before participating in the experiment.

Design. The design was essentially the same as in Experiment 1. The difference was the use of 3 experimental conditions : original (ORI), inverted (INV) and swapped (SWA). Every image was seen 2 times for each of the 3 conditions by each participant, as a target and as distractor when the task is reversed (resulting in 100 images x 2 repetitions x 3 conditions x 2 tasks = 1200 trials per participant).

Image manipulation. In the ORI condition, there was no manipulation of the images in the Fourier domain. In the INV condition, the amplitude spectra were switched between the two images presented on every trial. Thus, face images had the amplitude from vehicles, and vehicles from faces. In the SWA condition, on every trial, the amplitude of each image was replaced by an amplitude randomly selected among the 99 other exemplars of the category (each individual amplitude is used only once per run). For example, a face had the amplitude of another face, resulting in images with amplitude and phase from the same category but not from the same image. Examples can be seen in Figure 5.

Figure 5: Images used in Experiment 2. (a) *Principles of the different possibles combinations between amplitude and phase spectrums with examples for an image.* ***(b)*** *Examples of images used in the 3 conditions, faces on top and vehicles on bottom. It is clear from the observation of INVERTED and SWAPPED images that they are both very noisy.*

Results

A global look at the results reveals that the performance in Experiment 2 for the original (ORI) condition was somewhat better than in Experiment 1 despite the fact that the images used were exactly the same (see Figure 6). Much of this difference may well be caused by the change in background color (from black to gray), although inter-subject variability may also be involved. A first analysis of the global effects, using a 2-way ANOVA with factors Task (face or vehicle) and Type (ORI, INV, SWA), showed an effect of both on accuracy and mean RT (Type: $F_{(2,22)}=13.56$, $p<.001$; Task $F_{(1,11)}=44.80$, $p<.001$) and accuracy (Type: $F_{(2,22)}=84.06$, $p<.001$; Task: $F_{(1,11)}=43.56$, $p<.001$).

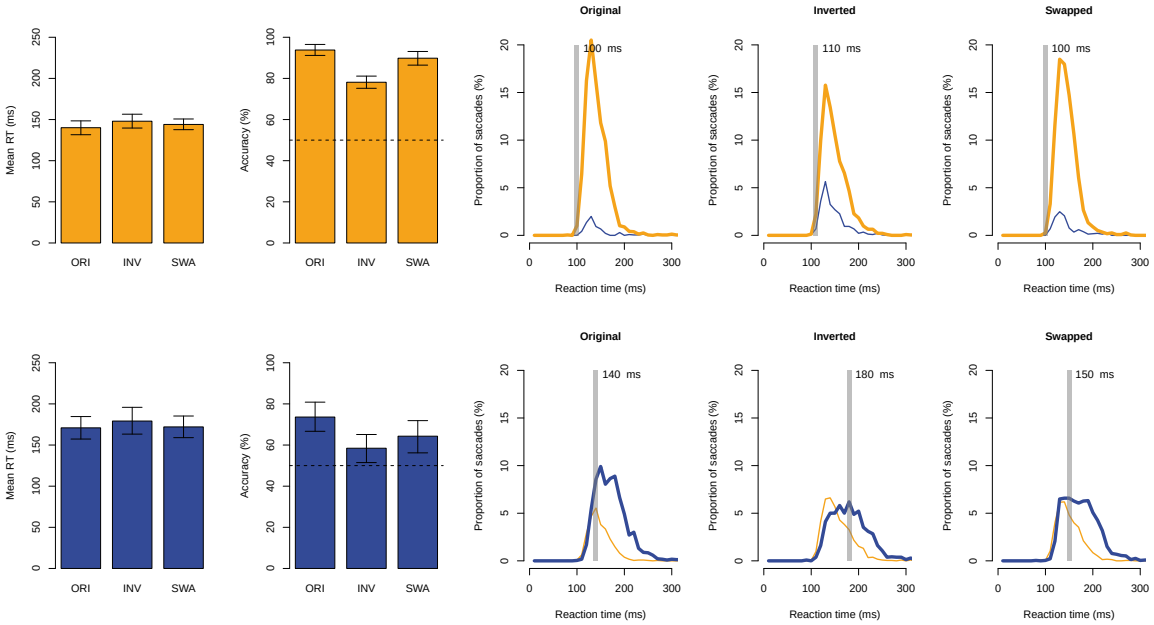

Figure 6: Results of Experiment 2. *Top: when participants had to saccade toward faces. Bottom: when participants had to saccade toward vehicles.* ***(a)*** *Mean RT and accuracy across subjects in the original (ORI), inverted (INV) and swapped (SWA) conditions. Error bars represent bootstrapped 95% C.I. of the mean* ***(b)*** *Distributions of RT in each condition and each task. Thick lines are for correct responses, thin lines for errors. Orange lines for saccades toward faces, blue lines for saccades toward vehicles. The transparent gray bar is placed according to the minimum RT (the first 10 ms bin of time where correct responses become significantly higher than incorrect ones).*

However, the principal aim of this experiment was to test the difference between the ORI condition and the two conditions involving image manipulations: INV and SWA. A post-hoc analysis using correction for multiple comparisons (Tukey HSD) was used, and revealed no effect on mean RT, thus only analysis of accuracy will be developed. In the face task, ORI (accuracy = 93.8%) and SWA (89.8%) leaded to comparable levels of performance and were significantly better performed than INV (78.1%). A closer look at RT distributions reveals that even though accuracy was lower in the INV condition, the first selective responses still occur very early. In the vehicle task, the effect is less clear, ORI (73.6%) and SWA (64.3%) are again comparable, ORI being the only condition different from INV (58.5%).

Each image category can be divided into two different object size: close-up and middle view. It could have been hypothesized that size would have an effect here, caused by the specific image manipulations. However, remarkably, a post-hoc analysis showed

absolutely no effect: there is no speed or accuracy differences between all the different conditions of size mixing. For example, we found no evidence that mixing the phase information from a close-up face with the amplitude information from a mid-distance view (or the inverse) specifically impaired performance in comparison to other conditions.

As a global conclusion of Experiment 2, the results argue in favor of a significant use of amplitude information by the visual system to guide saccades, and this use seems to be made independently of phase information. This strongly differs from what Gaspar and Rousselet reported using a manual response task, since congruency between phase and amplitude information does not seem to be crucial in the saccadic choice task, but rather can be used independently. Surprisingly, it seems that there is a tendency for amplitude information to also be used by subjects when the task was to target vehicles, even if it is less clear than for faces. Another surprise was that this effect seems to not be only observed for the fastest saccades. So even if, again, phase information seems to be the most important information, amplitude spectrum definitely plays a role in saccade generation. Intriguingly, their respective roles seem to be independent.

GENERAL DISCUSSION

Our first goal was to investigate the role of amplitude information in the Fourier spectrum in the generation of fast saccades toward faces. The question raised by this study follows from two recent results which showed that (i) saccades can be selectively oriented toward a target extremely fast (100-110 ms) in the specific case of this target being a face (Crouzet, et al., 2010), (ii) in the absence of phase information, amplitude alone can attract fast saccades by its own if it comes from a face image (Honey, et al., 2008). Our first aim was thus to test the influence of amplitude spectrum information in the effect observed in Crouzet et al., 2010 study. Experiment 1 showed that the absence of amplitude information tends to slow down saccadic responses. However, Gaspar and Rousselet recently demonstrated that this performance decrease could be explained, not by the lack of amplitude information, but rather by the edge noise added when normalizing this parameter. This lead to Experiment 2 which, using two new image manipulations, showed that subjects had remarkably good performance even if images are composed with phase and amplitude from the same object category but different images (and thus had a very high level of edge noise).

To sum up all the results. First, it seems that the bias toward faces over vehicles cannot be eliminated by the different manipulations made on amplitude spectrum. In other words, subjects were always better at saccading toward faces than toward vehicles, even when the amplitude spectrum was normalized or inverted. Thus, even though a face-like amplitude spectrum alone can attract fast saccades (Honey, et al., 2008), a large part of the bias is caused by phase related information. Second, information from the amplitude spectrum is effectively used by the visual saccadic system to drive saccades, and this use seems to be made independently of phase information.

Why amplitude spectrum was used here and not in most previous studies?

Several previous studies have claimed that amplitude spectrum information on its own is not important for natural scene recognition. Nevertheless, most of them used manual responses (Gaspar & Rousselet, 2009; Loschky & Larson, 2008; Loschky, et al., 2007; Wichmann, et al., 2006), which raises the possibility that the effects seen here might be specific to saccadic responses. However, a recent study had the same conclusion for an animal detection using a saccadic choice task similar to the one used here (F. A.

Wichmann, J. Drewes, P. Rosas, & K. R. Gegenfurtner, 2010). In the experiment 2 of this study, the authors first showed that animal detection is slightly impaired by amplitude spectrum normalization. Then, they used amplitude alone information to classify their images as animals or non-animals, and divided their set between images for which the classifier was the more confident (easy images) and those for which the classifier was the less confident (difficult images). Finally, with a post-hoc analysis, they showed that the accuracy differences between easy and difficult images was not only clear in the original condition, but also in the normalized one. The straightforward interpretation was that amplitude information had no causal role on their results for animal detection by human subjects. However, as can be seen in their table of results, their RTs were very slow compare to ours (most of their mean RT values were above 270 ms, whereas most of our mean RTs were below 200 ms). A contrasted analysis of their results between fast and slow reaction time would thus be of great interest (Honey, et al., 2008), and we would predict that the interpretation could be different according to this criterion. Thereby, the difference in reaction time can well explain the difference between the two studies, and thus suggests that the importance of amplitude information would be particularly critical early on during processing.

Underlying brain mechanisms

The results also show clearly the relative independence of amplitude and phase processing for face detection. This pattern could be the manifestation of two independent streams of processing in the brain: (i) a fast sub-cortical route involving the superior colliculi, pulvinar and amygdala, mostly driven by amplitude information and which could also feed the ventral stream (Johnson, 2005), (ii) the classic ventral stream from V1 to V4 and IT mostly driven by phase information. According to this hypothesis, both pathways would work in parallel and feed decisional units (probably in LIP or FEF) that finally initiate a saccade toward the target. Obviously, the existence of only two pathways is a simplification, since many others could exist (magnocellular vs. parvocellular, via the dorsal stream, etc).

Eccentricity and summary statistics

Time is not the only factor that has to be considered to explain the difference between manual and saccadic studies. In most manual tasks, the presentation of the images

was foveal (Gaspar & Rousselet, 2009; Wichmann, et al., 2006), whereas the presentations are peripheral in saccadic tasks (Honey, et al., 2008, as well as the present study). Could this difference accounts for the use of amplitude spectrum information in saccadic studies? It is well-known that contour, as well as positional information (Greenwood, Bex, & Dakin, 2009; Levi, 2008) is less and less precise with increasing eccentricity. Indeed, peripheral vision is characterized by a high degree of spatial uncertainty (D. G. Pelli, 1985). A result of this could be that patterns in a constrained region of the periphery would be processed as textures, so that individual patterns are not available for discrimination (Orbach & Wilson, 1999). This "jumbled" effect on contours would thus disrupt phase more than amplitude (unlocalized) information. In this case, the amplitude spectrum could be used as a "summary statistic" to detect objects in the visual field taking account of the low capacity of the visual system in the periphery. The relatively modest eccentricity we used in this saccadic choice task (8° from fixation to stimuli centers but image covering 14°x14° each of visual angle) could mean that there is still a role for phase information. However, at even higher eccentricities the amplitude spectrum may become even more important. To conclude, the two factors (slow vs. fast response, central vs. peripheral presentations) have not been completely dissociated here. Further investigations will definitely be needed to disentangle the contribution of time and eccentricity here.

Our conclusion is thus that, contrary to most studies that have used longer RT responses, information based on the amplitude spectrum alone can also be used by the visual system to detect objects in the visual field and produce extremely fast behavioral responses. More than a global Fourier analysis of the entire visual scene (Torralba & Oliva, 2003), a process which is rather unlikely in the real world, and is absolutely non-informative about the localization of objects, a more plausible mechanism would be the fast extraction of amplitude spectrum information in a localized fashion. This could take a patch-wise and multi-scale form, as in a wavelet analysis or in the Spatial Envelope model (Oliva & Torralba, 2006). This amplitude-based but localized information seems well-suited to be the basis of ultra-rapid processing of objects, as can be observed with a saccadic choice task. Further investigations will be needed to determine how exactly this type of information, present in the early visual system, is used to guide eye movement.

CONCLUSION

As a summary, our results are a clear demonstration of the use of amplitude spectrum information by the visual system. The use of this kind of information could be especially important, because it can be extracted extremely fast and thus allows to produce rapid selective responses to important stimuli in our visual field, even before a more precise analysis has been finished.

Acknowledgments - We would like to thank Guillaume Rousselet for his useful comments on the original idea and design of Experiment 2. This work was supported by the Délégation Générale pour l'Armement (DGA), the Fondation pour la Recherche Médicale (FRM) and the Agence Nationale pour la Recherche (ANR).

REFERENCES

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

Campbell, F. W., & Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *J Physiol, 197*(3), 551-566.

Crouzet, S., Kirchner, H., & Thorpe, S. (2010). Fast saccades towards face: Face detection in just 100 ms. *Journal of Vision*.

Dakin, S., & Watt, R. (2009). Biological "bar codes" in human faces. *Journal of Vision, 9*(4), 2.

Field, D. J. (1999). Wavelets, Vision and the Statistics of Natural Scenes. *Philos Trans R Soc Lond A, 357*, 2527-2542.

Gaspar, C. M., & Rousselet, G. A. (2009). How do amplitude spectra influence rapid animal detection? *Vision Res, 49*(24), 3001-3012.

Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2009). Positional averaging explains crowding with letter-like stimuli. *Proc Natl Acad Sci U S A, 106*(31), 13130-13135.

Guyader, N., Chauvin, A., Peyrin, C., Herault, J., & Marendaz, C. (2004). Image phase or amplitude? Rapid scene categorization is an amplitude-based process. *C R Biol, 327*(4), 313-318.

Hershler, O., & Hochstein, S. (2006). With a careful look: Still no low-level confound to face pop-out. *Vision Res, 46*(18), 3028-3035.

Honey, C., Kirchner, H., & VanRullen, R. (2008). Faces in the cloud: Fourier power spectrum biases ultrarapid face detection. *J Vis, 8*(12), 1-13.

Johnson, M. H. (2005). Subcortical face processing. *Nat Rev Neurosci, 6*(10), 787-798.

Joubert, O. R., Rousselet, G. A., Fabre-Thorpe, M., & Fize, D. (2009). Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *J Vis, 9*(1), 2 1-16.

Kaping, D., Tzvetanov, T., & Treue, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cogn, 15*(1), 12-19.

Keil, M. S. (2008). Does face image statistics predict a preferred spatial frequency for human face processing? *Proc R Soc Lond B Biol Sci, 275*(1647), 2095-2100.

Keil, M. S. (2009). "I look in your eyes, honey": internal face features induce spatial frequency preference for human face processing. *PLoS Comput Biol, 5*(3), e1000329.

Keil, M. S., Lapedriza, A., Masip, D., & Vitria, J. (2008). Preferred spatial frequencies for human face processing are associated with optimal class discrimination in the machine. *PLoS ONE, 3*(7), e2590.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Res, 46*(11), 1762-1776.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci, 23*(11), 571-579.

Levi, D. M. (2008). Crowding--an essential bottleneck for object recognition: a mini-review. *Vision Res, 48*(5), 635-654.

Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron, 62*(2), 281-290.

Loschky, L. C., & Larson, A. M. (2008). Localized information is necessary for scene categorization, including the Natural/Man-made distinction. *J Vis, 8*(1), 4 1-9.

Loschky, L. C., Sethi, A., Simons, D. J., Pydimarri, T. N., Ochs, D., & Corbeille, J. L. (2007). The importance of information localization in scene gist recognition. *J Exp Psychol Hum Percept Perform, 33*(6), 1431-1450.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*: Henry Holt and Co., Inc. New York, NY, USA.

Nestor, A., Vettel, J. M., & Tarr, M. J. (2008). Task-specific codes for face recognition: how they shape the neural representation of features for detection and individuation. *PLoS ONE, 3* (12), e3978.

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision, 42*(3), 145-175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Prog Brain Res, 155*, 23-36.

Oppenheim, A. V., & Lim, J. S. (1981). The Importance of Phase in Signals. *Proceedings of the Ieee, 69*(5), 529-541.

Orbach, H. S., & Wilson, H. R. (1999). Factors limiting peripheral pattern discrimination. *Spat Vis, 12*(1), 83-106.

Pelli, D. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision, 10*(4), 437-442.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *J Opt Soc Am A, 2*(9), 1508-1532.

Piotrowski, L. N., & Campbell, F. W. (1982). A demonstration of the visual importance and flexibility of spatial- frequency amplitude and phase. *Perception, 11*(3), 337-346.

Potter, M. C. (1975). Meaning in visual search. *Science, 187*(4180), 965-966.

Smeets, J. B., & Hooge, I. T. (2003). Nature of variability in saccades. *J Neurophysiol, 90*(1), 12-20.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu Rev Neurosci, 19*, 109-139.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network, 14*(3), 391-412.

Vanrullen, R. (2006). On second glance: Still no high-level pop-out effect for faces. *Vision Res, 46* (18), 3017-3027.

Westheimer, G. (2001). The Fourier theory of vision. *Perception, 30*(5), 531-541.

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Res, 46*(8-9), 1520-1529.

Wichmann, F. A., Drewes, J., Rosas, P., & Gegenfurtner, K. R. (2010). Animal detection in natural scenes: Critical features revisited. *Journal of Vision, 10*(4), 1-27.

Wichmann, F. A., Drewes, J., Rosas, P., & Gegenfurtner, K. R. (2010). Animal detection in natural scenes: Critical features revisited. *Journal of Vision*.

2.4 Questions en suspens

Nous avons donc vu que le système visuel était capable d'extraire en un temps record une information lui permettant de détecter la présence d'un visage dans le champ visuel. Nous avons ensuite montré que l'information contenue dans le spectre d'amplitude pourrait, au moins en partie, lui servir d'intermédiaire pour réaliser cette tâche si rapidement. De nombreuses études devront encore être menées pour réellement découvrir le(s) secret(s) de cette détection.

Dans une série de nouvelles études, nous avons par exemple commencé à tester ses limites au niveau spatial. Jusque là, la tâche de choix saccadique consistait en un simple choix droite/gauche et les images utilisées jusqu'ici contenaient clairement un visage ayant une taille assez importante. Afin de déterminer jusqu'où peut aller ce système, nous avons donc récemment réalisé plusieurs expériences afin de tester les effets de taille, ainsi que la précision spatiale des saccades orientées vers ces visages en fonction de leur latence. Nous espérons que ces expériences nous permettront d'aller plus loin dans la compréhension des mécanismes de détection d'objets [2] [3]. Une autre voie consisterait aussi à utiliser des images de visages contenant différents types d'informations : par exemple, tester si ces saccades ultra-rapides résistent lorsque les visages sont des visages de cartoon, de caricature, lorsque la polarité est inversée, etc.

De plus, comme il a été souligné dans la discussion de l'article 2, l'impact de l'excentricité dans l'utilisation des informations de spectre d'amplitude reste à déterminer précisément et pourrait s'avérer très intéressant. Celui-ci pourrait en effet être un candidat pour tenir le rôle de statistiques résumées (anglais : summary statistics) dans la périphérie du champ visuel. De plus en plus d'études s'intéressent explicitement à la question de la représentation utilisée par le système visuel pour coder les objets périphériques (Levi, 2008; Pelli et Tillman, 2008; Pelli, 2008; Alvarez et Oliva, 2009, 2008). Les saccades sont certainement un outil de choix pour étudier cette question car leur contrôle est, par essence, basé sur ce type d'information.

2. Mathey, M. A., Crouzet, S. M. & Thorpe, S. J. (2010) Ultra-rapid saccades to faces : the effect of target size. VSS, Naples, Florida.

3. Marie Mathey, Sébastien M. Crouzet & Simon J. Thorpe (soumis) The accuracy of ultra-rapid saccades to faces. ECVP 2010, Lausanne, Switzerland.

Chapitre 3

Contenu des traitements précoces

Sommaire

Le protocole de choix saccadique permet donc, en ouvrant une fenêtre sur un état précoce de l'information, de mettre en évidence des effets qui n'apparaissaient pas avec une réponse manuelle plus tardive. Cet outil s'avère donc extrêmement intéressant pour l'étude des mécanismes de reconnaissance rapide d'objets en général. Dans les deux études que je présente dans ce chapitre, nous avons essayé de préciser la nature des traitements mis en jeu. Dans un premier temps, nous verrons l'influence des informations contextuelles sur la reconnaissance d'objets (une vache dans un pré est-elle reconnue plus rapidement qu'une vache dans un bureau ?). En effet, les images présentées étant toujours des scènes naturelles, elles ne contiennent pas seulement l'objet en lui-même. Le système visuel pourrait donc exploiter les informations contenues dans le reste de l'image pour aider la reconnaissance d'objet. Dans la deuxième étude, nous avons voulu déterminer la précision des traitements permettant de guider les saccades. Nous avions pour l'instant démontré les très bonnes performances des sujets pour faire des saccades sélectives vers les catégories que sont *animal*, *visage* ou *véhicule*. Qu'en

est-il si la cible correspond à une catégorie plus précise ? Par exemple, on peut demander aux sujets de faire des saccades sélectivement vers les chiens, alors que les distracteurs sont d'autres animaux. En d'autres termes, cette étude s'intéressera à l'effet du niveau de catégorisation (voit-on un chien d'abord comme un chien ou comme un animal ?). L'objectif de ces deux études est que le protocole de choix saccadique nous donne accès à des étapes précoces des traitements visuels, permettant ainsi d'étudier la dynamique de ces effets de façon plus précise que ce qui était permis jusqu'ici par les réponses manuelles.

3.1 Modulation par le contexte

Malgré leur complexité, les scènes naturelles sont hautement structurées et de nombreuses redondances peuvent en être extraites. Sans entrer dans des statistiques complexes et en restant dans le domaine de l'image, on peut déjà déterminer certaines régularités. Par exemple la plupart des objets sont plus susceptibles d'apparaître dans certains contextes que dans d'autres (par exemple une vache dans un pré/autoroute, une armoire dans une maison/train). Il est donc plus que probable que le système visuel utilise ces régularités pour optimiser son fonctionnement et même générer des prédictions (Bar, 2007).

Concernant la reconnaissance rapide d'objets, le système visuel peut-il utiliser ces associations pour optimiser la reconnaissance ? Dans le modèle de guidage contextuel de Torralba par exemple, l'analyse préalable du contexte permet de contraindre le processus de localisation d'objet (Torralba *et al.* 2006, voir section 1.5 pour plus de détail). Cependant, un tel rôle pour le contexte implique qu'il doit être traité plus rapidement que l'objet. Cette question a été abordée par une étude de notre équipe (Joubert *et al.*, 2007). Ils ont d'abord montré que les sujets pouvaient réaliser rapidement et avec un très bon taux de réussite une tâche go/no-go où ils devaient catégoriser les images comme *naturelles* (mer, montagne, plage, etc) ou *manufacturées* (villes, immeubles, etc). La superposition parfaite des courbes de temps de réaction obtenues pour la catégorisation de contexte avec celles obtenues dans une tâche animal/non-animal (Rousselet *et al.*, 2003a) suggère une dynamique similaires entres les traitements de l'objet et du contexte global. Il parait donc difficile dans ce cadre d'imaginer un processus tel que celui du modèle de guidage contextuel, en tout cas pour les réponses les plus rapides. Par contre, des interactions entre les deux processus sont tout à fait envisageables dans ce cadre. Afin de préciser les interactions entre les traitements de l'objet et du contexte, ils ont donc réalisé une nouvelle étude basée sur une tâche animal/non-animal (Joubert *et al.*, 2008). La particularité de cette étude consistait à avoir découpé précisément les images d'animaux pour les coller ensuite sur deux types de contextes (naturel ou manufacturé). De cette façon, ils avaient à leur disposition un groupe d'images bien contrôlées où une

même image d'objet pouvaient apparaître soit dans un contexte *congruent* (animal sur fond naturel), soit *incongruent* (animal sur fond manufacturé). Les résultats montrent un avantage des essais congruents sur les essais incongruents : les sujets étaient plus rapides et plus précis pour détecter la présence d'un animal sur fond naturel que sur fond manufacturé. Surtout, cet effet était présent dès les toutes premières réponses des sujets. Il semble donc que, plus qu'une facilitation du traitement de l'objet par le contexte comme celle du modèle de Torralba, les mécanismes mis en œuvre ici soient de nature purement ascendante. L'hypothèse est que l'objet et le contexte seraient traités en parallèle, et que des interactions directes apparaissent dès le départ (et dans les deux sens, Davenport et Potter, 2004).

Comme nous l'avons vu précédemment, les expériences utilisant une tâche de choix saccadique ont montré que la détection d'objets pouvait être réalisée encore plus vite que ce que les études utilisant des réponses manuelles avaient montré jusqu'ici. Ce protocole pourrait donc nous permettre d'aller observer si les interactions entre les traitements de l'objet et du contexte ont bien lieu dès les toutes premières réponses, ou si dans une fenêtre temporelle précoce, ils se font indépendamment.

3.1.1 Résumé de l'étude

La congruence du contexte peut biaiser la catégorisation rapide d'objets dès les toutes premières réponses comportementales lors d'une tâche go/no-go de réponse manuelle. Le but de cette étude consiste à déterminer les latences les plus courtes auxquelles un tel effet de congruence peut apparaitre. À partir d'une tâche de choix saccadique, les sujets ont participé à une discrimination d'objets (animal vs. véhicule) et de contexte (naturel vs. manufacturé). L'ordre des catégories cibles et des conditions de congruence étaient contrebalancées sur l'ensemble des sujets. Cette expérience a permis de montrer que :

- Les sujets sont plus rapides et ont un meilleur taux de réussite pour faire des saccades vers les animaux que vers les véhicules (un biais comparable même si moins fort qu'entre les visages et ces mêmes véhicules), un résultat clairement en opposition avec les résultats obtenus précédemment à l'aide d'un protocole go/no-go (VanRullen et Thorpe, 2001).
- Les contextes naturels et manufacturés peuvent être discriminés en seulement 160 ms, ceci prend plus de temps que lorsque la cible est un animal, mais autant que pour les véhicules (avec un taux de réussite clairement meilleur que pour les véhicules).
- L'effet de facilitation du contexte congruent sur la détection de l'animal est retrouvé. Cependant, et de façon très intéressante, il n'est pas présent dès les premières réponses, et n'apparait qu'environ 160 ms après l'affichage des images.

Le biais vers les animaux fait donc directement écho au biais vers les visages que nous avions mis en évidence dans l'article 1 (section 2.2). Même si celui-ci semble moins puissant, il présente les mêmes caractéristiques : asymétrie dans les temps de réaction, précision très basse lorsque la saccade doit être initiée vers le véhicule, et même la tendance, pour les saccades les plus rapides, à être plus souvent vers l'animal (ici distracteur) que vers la cible véhicule. Ce phénomène peut donc aussi expliquer pourquoi la précision pour les véhicules se retrouve si basse (plus basse que pour les contextes). Ce qui n'était pas le cas dans l'expérience 1 de l'article 1 où les distracteurs utilisés étaient "neutres". Dans cette expérience, les véhicules étaient traités plus lentement que les visages et les animaux, mais avec un niveau de précision bien plus haut que ce que nous observons ici. On peut donc imaginer que les distracteurs animaux ont clairement perturbé le traitement des véhicules.

Partant de l'observation que les distracteurs animaux ont fortement biaisé la tâche lorsque les sujets devaient faire des saccades vers les véhicules, on peut imaginer que le traitement des véhicules s'avère au final très similaire à ce que l'on observe pour les contextes, avec des premières saccades sélectives plus tardives que pour les catégories "favorites" que sont animal et visage. Ceci pourrait donc correspondre au temps "normal" requis pour sélectionner l'information pertinente et initier une saccade. Ce temps "normal" ne s'appliquerait pas à certains types d'objets, ceux pour lesquelles des mécanismes spéciaux seraient déjà implémentés dans le système visuel afin de les optimiser. Dans ce cadre, et même si les informations diagnostiques pouvant servir pour le contexte sont facilement et rapidement extractibles (spectre d'amplitude de Fourier par exemple), les premières réponses vers les animaux restent non-influencées par ces informations contextuelles. Cependant, comme le montre clairement nos résultats, dès que celles-ci sont disponibles (ici autour de 160 ms), elle influencent alors très clairement la reconnaissance de l'animal.

3.1.2 Article 3 : Timing the earliest object-context interactions

Timing the earliest object-context interactions

(en préparation)

*Olivier R. Joubert a,b

*Sebastien M. Crouzet a,b

Simon J. Thorpe a,b, & Michele Fabre-Thorpe a,b¤

a Université de Toulouse, CerCo, UPS

b CNRS, UMR 5549, Faculté de Médecine de Rangueil, Toulouse, France

* The two authors equally work on experimental design result analysis and publication

¤ corresponding author

Centre de Recherche Cerveau et Cognition, CerCo, UMR 5549, Faculté de Médecine de Rangueil, 133 route de Narbonne 31062 Toulouse, cedex 9 France

Email : michele.fabre-thorpe@cerco.ups-tlse.fr

Abstract

Background. Rapid categorization of animals in 20 ms flashed natural scenes is faster and more accurate when surrounding context is congruent, even in the case of the fastest manual responses. What is the earliest latency at which context can influence object categorization?

Results. We used a saccadic choice task in which target and distractor are simultaneously flashed in the left and right visual fields and subjects must saccade as fast as possible towards the target-side. Because of short saccade latencies, this recently developed task allows to investigate a previously inaccessible temporal window of processing. We first showed a better performance of subjects at targeting animals than vehicles (81% vs. 63% correct, 120 ms vs. 180 ms minimal saccade latency). Using "context" categories as targets such as "natural vs. man-made scenes", the task is performed with a higher accuracy (72-73%) than with vehicle targets, but with a shorter minimal saccade latency (160 ms). Finally, using congruent vs. incongruent object/context associations, we demonstrated that the processing of animal targets was affected by context congruency only for saccades with latency over 160 ms, a delay that corresponds to the earliest correct saccades observed when using context categories as targets. Such interactions were not seen with vehicles, a result that might reflect the fact that vehicles can be congruent in both types of contexts.

Conclusions. These results reveal new strong constraints on the ascending processing of context and object/context interactions time-courses that biological and computational models should take into account.

Introduction

Evolution and experience have shaped our visual system for dealing efficiently with the surrounding world. Indeed, although natural scenes are very rich and complex, they are also meaningful and highly structured. Certain types of object are more likely to occur in particular contexts. Thus, wild animals are more likely to be seen in natural landscapes, and cars are more likely to be seen in a street scene. It seems reasonable to suppose that the visual system could take advantage of these object/context regularities to generate predictions, anticipate the relevant future and provide adequate responses effortlessly [1,2]. This type of implicit knowledge can be used for example to facilitate visual object recognition via top down influence of contextual information [3-5].

In many cases these top-down influences can be set up before a scene has been presented. However, several studies have shown that both object and context categories can be extracted in a fraction of a second from a flashed scene suggesting that there is a temporal window during which context could influence object processing in the ascending processing of an individual scene. Among the first such studies, Potter [6] revealed the remarkable ability of our visual system to grasp the meaning of scenes embedded in a RSVP (Rapid Serial Visual Presentation) sequence at 8 images per second. Following the study by Thorpe et al. [7] that used a go/no-go rapid categorization task, our group has shown the remarkable efficiency with which the human visual system can detect the presence of an animal, a vehicle, a human face or an animal face within natural scenes flashed for only 20 ms [8-12]. In these studies, subjects were very efficient, typically scoring over 94% correct with median reaction time (RT) of about 400ms. The earliest responses (excluding anticipations) could sometimes be observed as early as 250 ms, suggesting purely feed-forward visual processing [7,10,13]. Humans can thus categorize natural scenes on the basis of whether or not they contain a target-object from a given category, but they can also categorize the global contextual frame (natural vs. man-made) of the flashed scene with an accuracy of about 90% and median RT below 400 ms [14,15]. Such rapid access to scene gist could rely on the extraction of 3D primitives [16], as well as global statistics or/and spatial layout [17-23].

Processed in the same temporal window, object and scene processing could interact, and indeed, reciprocal bottom-up interactions have recently been demonstrated using natural photographs. The presence of a salient object has been reported to interfere with the

processing of scene context especially when incongruent with the scene [14], and object processing is more accurate and faster in congruent scenes when compared to incongruent scenes [24, 25]. Moreover, such bottom-up object/context interactions develop very early as they can be observed even for the earliest responses made by the subject, responses that occur around 250-270ms.

To explain the influence of context on object processing, a model has been proposed by Bar and colleagues in which blurred, low spatial-frequency representations of the stimulus are processed rapidly to extract the likely contextual frame, a role thought to be played by the parahippocampal cortex, and to suggest potential objects compatible with these coarse representations, a role suggested to involve prefrontal cortex. The inferotemporal cortex would be fed both types of information and would activate the representation of the object most likely to be present in the given contextual frame [26,27]. Early activity linked with object processing has been seen in the orbitofrontal cortex (OFC, [26]) as well as in the frontal eye fields in relation with the sensory characteristics of visual stimuli [28]. On the other hand, although object category information can be decoded from human visual cortex as early as 100 ms poststimulus [29], the influence of OFC on the fusiform gyrus was not seen until slightly later, around 130 ms after stimulus onset [30]. An alternative interpretation of the very early contextual effects reported by Joubert et al. [25] is based on the processing of the low spatial-frequency representations of both the contextual frame and the target object within the magnocellular pathway of the ventral visual stream. Facilitation would take place when the selective populations of neurons that are activated in higher visual areas such as the infero-temporal cortex have frequently been co-activated through experience. In contrast, interference would take place when co-activations of such populations are rarely seen or even conflictual.

The study by Joubert et al [25] showed that even the earliest manual responses to animal-targets were influenced by the surrounding context, implying that the effect had already taken place before 250-270 ms. But such data cannot tell us precisely when the effect had taken place. The aim of the present study was therefore to explore object processing, context processing and their interactions in an even earlier temporal window by using a saccadic choice task developed by Kirchner and Thorpe [31]. The task takes advantage of the fact that two images presented side by side can be processed in parallel with essentially no behavioural cost [32, 33]. In this new protocol, subjects are shown two scenes for 400ms and are required to make a saccade as quickly and accurately as possible towards the scene that contained a given target such as an animal or a human face. Tested in such task, humans are able to initiate a saccade with latencies that can be as short as 120-130 ms for animal targets [31] and even shorter (100 ms) for face targets

[34]. Studies using backward masking showed that the processing dynamics are similar to those seen with the manual go/no-go task [35]. To account for such short latencies in this forced-choice task, the authors suggested that visual information processed in early visual areas up to V4 may be sufficient.

In the present study, we used this saccadic choice task to evaluate and compare the temporal dynamics of object and context categorization as well as object/context interactions under these very severe temporal constraints. By using as targets two context categories (“Man-made” and “natural”) and two object categories (“Animal” and “Vehicle”), we evaluated how long the visual system needs to extract enough information for performing a context discrimination task, and whether object processing in this very early temporal window can already be modulated by context congruency.

Methods

Participants

12 participants all volunteered, (9 females, mean age 21, range 18-26, 10 of them right-handed) gave their written informed consent and had normal or corrected-to-normal vision.

Stimuli

Altogether, 864 photographs of natural scenes were used in this experiment (see Figure 1) of which 96 were used for training and 768 in the testing blocks. The original images were selected from a large commercial CD-Rom library (Corel Stock Photo Libraries) or from the web. They were cropped to 400x400 pixels and converted to grey-level (8-bit jpeg format). Their global luminance and RMS contrast were then normalized by taking the average luminance and contrast of the 864 grey-level scenes. Stimuli were divided in 3 main sets (Figure 1): scenes that did not contain any foreground object (432), scenes containing one or more animals (216), and scenes containing one or more vehicles (216). Animals included mammals, birds, rodent, snakes, fishes and insects, while the vehicles included cars, trucks, planes, boats, bicycles, motorbikes, carriages, etc. A given scene could not contain both an animal and a vehicle. Each of the 3 sets was divided in two equal subsets depending on the global environment: natural or man-made. The natural environment images included photographs of coasts, mountains, fields, forests, deserts, icebergs, lakes and savannas, while the man-made environments included photographs of streets, places and buildings from all around the world using outdoor, indoor views and

some aerial views. The congruence between the object and its context was defined in terms of “Man-made” vs. “Natural” terms, vehicles being man-made objects and animals natural objects. Among scenes with foreground objects, we thus considered animals in congruent natural context (Ani C), animals in non-congruent manmade context (Ani nC), vehicles in congruent manmade context (Vehi C) and vehicles in non-congruent natural context (Vehi nC). All stimuli were original contrary to previous studies in which stimuli were built by pasting objects on various backgrounds.

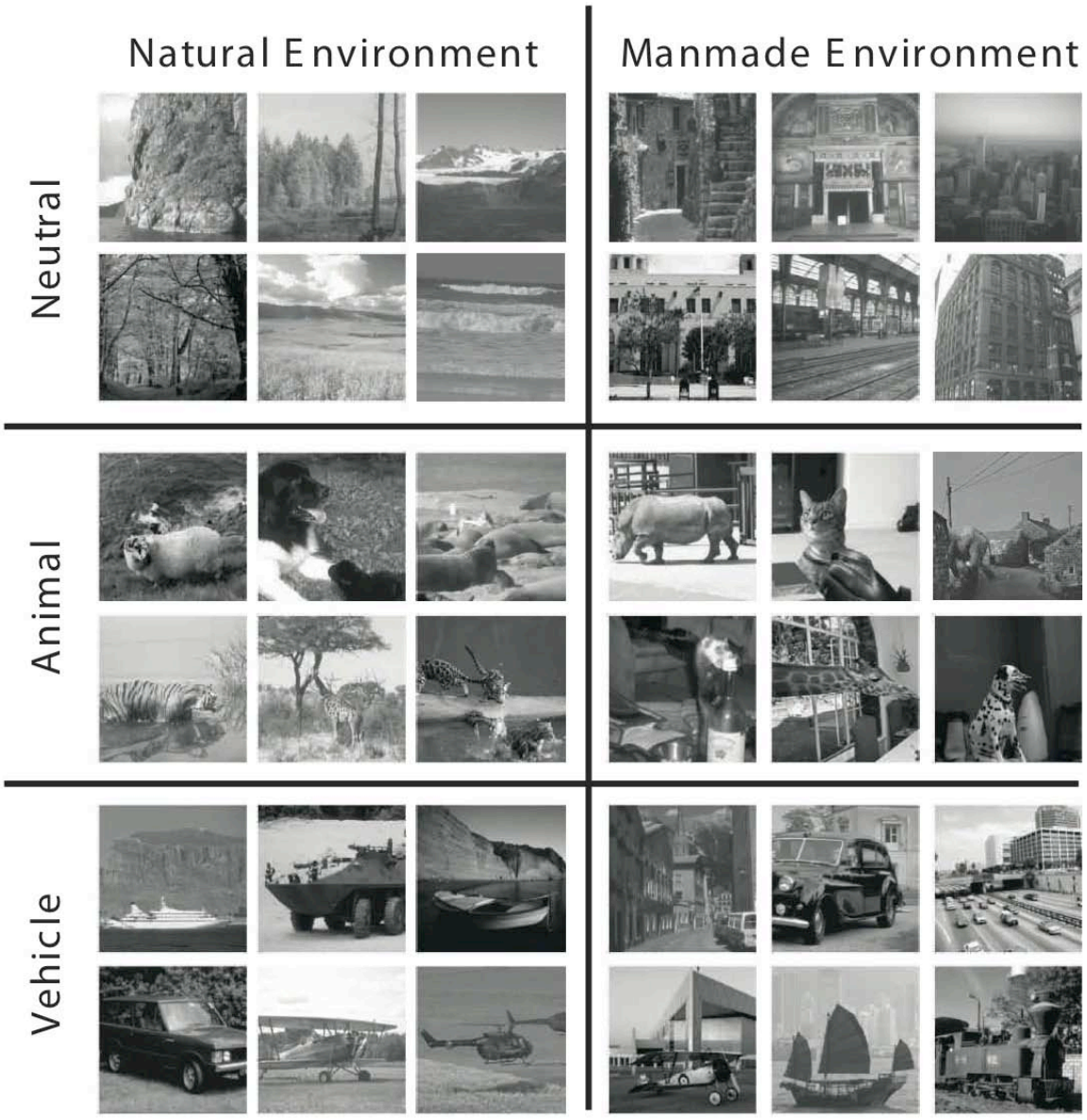

Figure 1: *Examples of stimuli used in this study. In the two contextual discrimination tasks, only neutral natural and man-made environments (without salient objects) were used as targets or distractors according to the task. In the two other object discrimination tasks, targets and distractors were animals or vehicles embedded in either natural or man-made scenes. Average luminance and RMS contrast were equalized across the set of stimuli.*

Task and Design

On each trial, after a fixation cross displayed for a pseudo-random time interval (800 – 1600 ms) to avoid anticipations, two natural scenes (one target and one distractor) were flashed for 400 ms in the left and right hemifields (see Figure 2.A.) centred at 8.6° of eccentricity. Subjects were asked to make a saccade as fast as possible to the side of the target. The trial ended by a blank screen displayed for 400 ms before the next trial started. A trial lasted between 1600 and 2400 ms. In this saccadic choice task, scenes were presented for much longer than in the manual task used by Joubert et al. [25] in which scenes were flashed for just 20 ms to prevent exploratory eye movements. But in this protocol, we were recording the first eye movement and the use of longer stimulus presentation times reduces the number of late saccades, resulting in a sharpening of the Saccadic Reaction Time (SRT) distribution [34].

Targets were equiprobable in both hemifields and their nature changed according to the instructions given at the beginning of each block. In all, the participants performed 4 blocks of the saccadic choice task, and each block included 4 runs of 48 trials preceded by 24 trials of training. During the experiment, each subject performed 2 blocks of context discrimination in which the scenes without foreground objects were used. In one of the blocks the targets were natural scenes whereas in the other block, the targets were man-made scenes. The two remaining blocks required subjects to target a given object category regardless of its context. One block used animal targets vs. vehicle distractors, while the other used vehicle targets vs. animal distractors. In these testing blocks, 50% of the targets and distractors were embedded in a man-made scene, the other half in a natural scene. Each picture was seen twice by each subject, once as target and once as distractor and the pairing of the target and distractor stimuli was random. Task order and presentation hemifield of the stimuli were counterbalanced across subjects.

Apparatus

Participants sat in a dimly lit room with their head position constrained by a chin rest and a forehead support. Grey-scale photographs were presented on a computer screen (19”, resolution 1024 x 768, vertical refresh: 100 Hz) on a mean grey background at a distance of 60 cm resulting in an image size of 14° x 14.5° and a horizontal eccentricity of 8.6°. Image presentation was carried out using Matlab and the Psychophysics Toolbox 3 [36,37]. Eye movements were monitored with an IView Hi-Speed eyetracker (SensoMotoric Instruments, Berlin, Germany). This infrared tracking system samples eye

position at 240 Hz. Saccade detection was performed off-line using the saccade based algorithm of the SMI BeGaze Event Detection package [38]. Before each run, a 13-point calibration was performed. We considered as saccadic reaction times (SRT) the latency of the initiation of the first saccadic response if entering one of the 2 images. SRTs below 80 ms and over 800 ms, if any, were discarded.

Statistics

To statistically estimate how significant was the performance difference observed between two conditions both in terms of global accuracy and median SRT, we used a bootstrap method based on Monte Carlo simulations using the following procedure. For each pair-wise comparison, the individual data for the 12 subjects in each of the two conditions was pooled together, randomly shuffled, and redistributed in fake samples with the same sizes as the original samples. The differences between these two new fake populations were then stored. This procedure was run 2000 times, providing a normal distribution based on the null hypothesis that the two conditions were actually sampled from the same population. The p-value was computed by evaluating the number of theoretical differences more extreme than the experimental one.

We also defined a "minimum SRT". Using a 10 ms time bin SRT distribution, minSRT corresponds to the first bin in which correct responses significantly outnumber errors using a $\chi 2$ test with a criterion of $p<.05$ when followed by 4 consecutive bins also reaching this criterion.

To compare speed of performance in congruent and non congruent conditions, we first computed for each of the two conditions, a single SRT distribution by substracting false alarm SRTs from hit SRTs. Then, to determine the earliest latency at which a congruency effect (if any) was observed, the two distributions were statistically compared using the same $\chi 2$ test procedure.

Results

Comparing Object and Context discrimination

Although the pairs of stimuli were displayed for 400 ms, subjects had to make a saccadic response towards the target scene as fast and as accurately as possible. When working with animal targets, the saccades performed by the subjects were correct in 80.9% of the trials and initiated with a median SRT of 181 ms (Figure 2B). Moreover, minSRT computed on the group of subjects showed that fastest subjects were able to trigger

saccades towards the correct side with SRTs as short as 120 ms (Figure 2C). These results are comparable with those obtained by Kirchner & Thorpe [31] using the same animal category. Such high levels of performance were not observed using vehicles as targets. Subjects were correct in only 63.2% of the trials with a median SRT of 207 ms and a minimal SRT of 180 ms (Figure 2 A, D). Monte Carlo simulations revealed that this bias between object categories in favor of animals was highly significant (accuracy: $p = 0.001$, median RT: $p < 0.0001$). When performing forced choice saccadic tasks with scene contexts as targets, subjects reached similar performance regardless of the target category (natural or man-made environments) with an accuracy rate of respectively 72.2% and 73.8% and a median SRT of respectively 217 and 212 ms (Figure 2A), values that did not differ significantly between the two contexts (accuracy: $p = 0.53$, median RT: $p = 0.47$). Furthermore, the MinSRT was 160 ms for both context categories (Figure 2 E-F). In the same experimental design, natural and man-made scenes without any foreground objects were thus discriminated less accurately and slower than animals (accuracy: both $p < 0.005$, median SRT: both $p < 0.0001$) but more accurately than vehicles (both $p < 0.005$) with similar median SRT (n.s.) and lower minSRT.

This pattern of results was also observed in the SRT distributions illustrating the processing time-course for each task. While the SRT distributions and accuracy for vehicle targets and for both contexts were similar (Figure 2 D-F), the SRT distribution for animal targets differed, with a sharper distribution and a peak that was shifted towards earlier latencies (Figure 2C). The response bias in favor of animals is reflected by the substantial number of early incorrect responses towards animals observed in the "vehicle" task (Figure 2D). When triggered early, there appears to be a bias for saccades to be spontaneously produced towards the scene that contains an animal even when the task requirement is to saccade towards "vehicles".

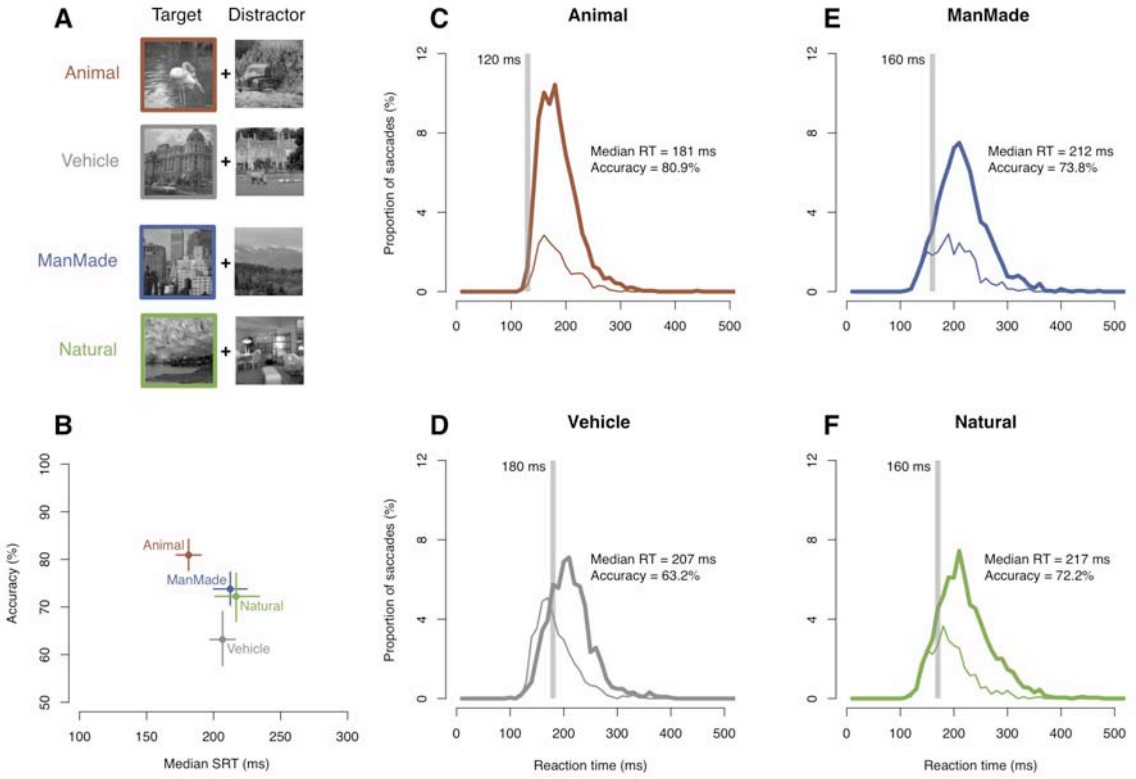

Figure 2: *Behavioral results. (A) example of pairs of stimuli in each of the 4 tasks indicating the color code used in B-F. (B) accuracy as a function of median SRT for each discrimination task. Horizontal and vertical bars correspond to 95% confidence interval for median RTs and accuracies respectively. The values are computed using a percentile bootstrap with 2000 resamples. (C-F) SRT distributions for correct (thick line) and incorrect (thin line) responses for each task with the percentage of responses pooled across all subjects and expressed over time using 10 ms time bins. While context processing time-course are very similar in both context tasks (E and F, respectively man-made and natural), minimal RT indicated by a grey thick vertical line was found at 160 msin both cases. (C and D), SRT distributions confirm the object category asymmetry : animal responses (C) are observed to be in average faster than vehicle ones (D) and with shorter minimal RT (respectively 120 and 180 ms). Note that saccades towards animals are difficult to control, since short latency saccades in the vehicle task are biased towards the scene that contained an animal (D).*

Context influences on object discrimination

The main aim of the study was to analyse whether object and context processing can interact at very short latencies and to determine the earliest latency for such interactions. Animal and vehicle targets were presented equally in natural or man-made contexts and context influences were computed on trials where both targets and distractors were congruent or incongruent. No difference was observed between congruent and

incongruent conditions in the vehicle task, correct saccades were produced with the same accuracy and the same speed (Figure 3B, respectively, 63.8% vs. 62.2%, n.s. and 210 ms vs. 203 ms, n.s.). Furthermore, the spontaneous attractiveness of animal targets was observed regardless of the context in which the animals and vehicles were embedded. Indeed, in the vehicle task (Figure 3D), early saccades are incorrectly produced towards animals even when both objects were presented in incongruent contexts. On the other hand, correct saccades towards animals were performed more accurately in the congruent vs. incongruent condition, albeit with similar median RTs (Figure 3.A., respectively, 87.5% vs. 75.5%, $p = 0.002$ and 186 ms vs. 178 ms, n.s.).This higher accuracy when animals are embedded in congruent contexts was also observed for the SRT distribution (Figure 3.C.) that had a higher peak of correct responses and a lower rate of incorrect responses in the congruent condition (vs. incongruent). To define the earliest latencies at which this context congruency effect occurs, we computed for each of the two conditions, a single SRT distribution by substracting false alarm SRTs from hit SRTs, we then compared the two congruent and non-congruent conditions using $\chi 2$ test on each 10 ms bin. While no difference was observed for vehicle task over the whole range of saccadic response time, animals were better categorized when embedded in congruent natural context and this effect was observed as early as 160 ms As indicated on Figure 3,C. Interestingly, this value happens to correspond to the min SRT obtained in the two context saccadic choice tasks.

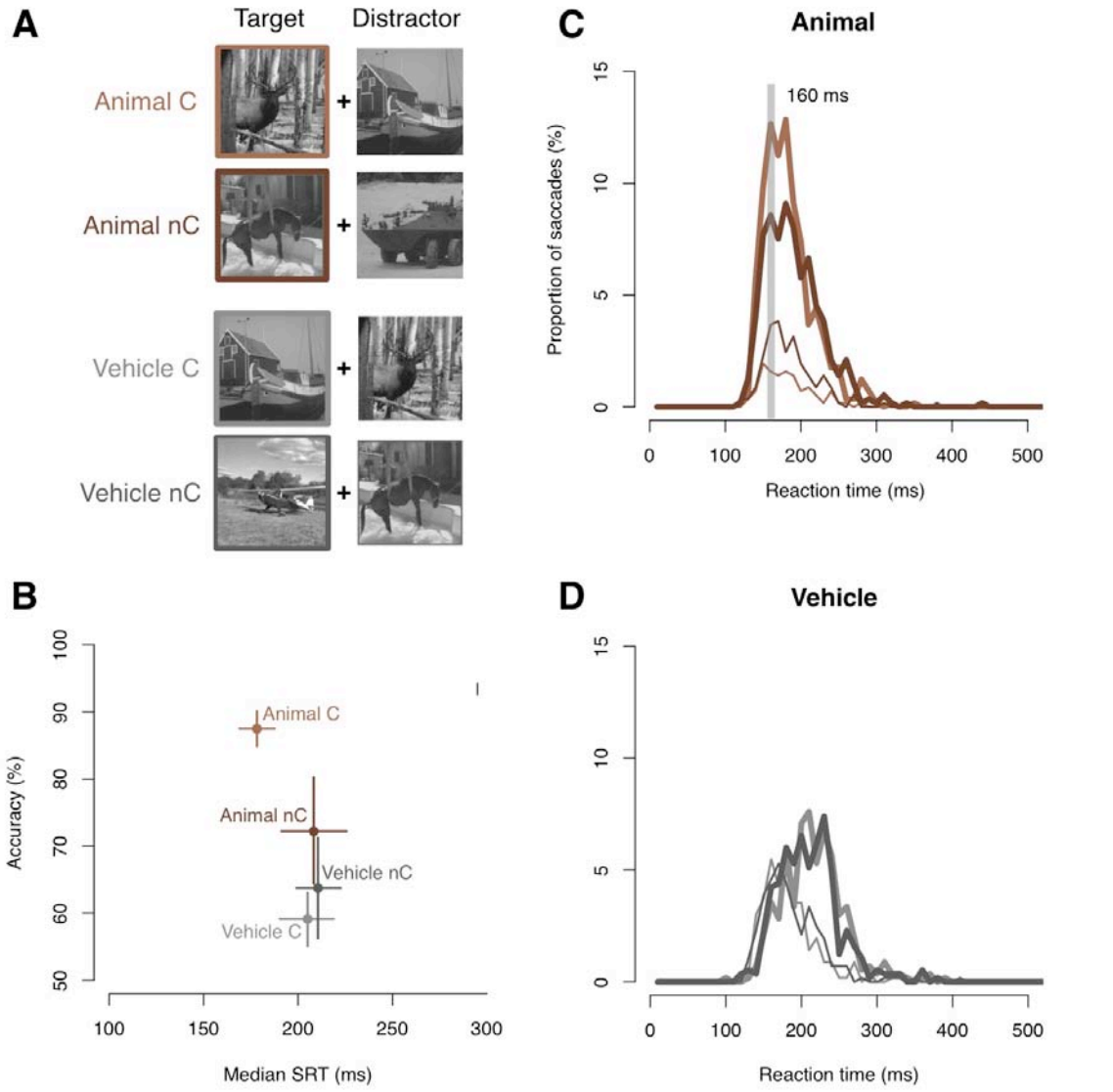

Figure 3: *Behavioral results in animal (brown) and vehicle (grey) tasks when both target and distractor were either congruent (C, light) or incongruent (nC, dark). (A) Examples of trials for each congruency condition in both object tasks. (B-D) Overall, results show a strong congruency effect in animal task and not in vehicle task. (B) Accuracy as a function of median SRT is very similar in both congruency conditions for vehicle tasks. In contrast, animals in congruent natural contexts are detected more accurately and faster than in incongruent manmade contexts. (C) SRT distributions in the animal task show a higher peak of correct responses (thick lines) and a lower peak of incorrect responses (thin lines) when the context is congruent (vs. incongruent). A $\chi 2$ test computed on Hits minus FA distributions revealed a difference between congruent and non-congruent responses starting at 160 ms indicated by the grey thick vertical line. (D) In the vehicle task, the SRT distributions are similar in both congruency conditions and reveal a bias towards incorrect animal responses for the shortest SRT of the distribution.*

Discussion

The first aim of the present study was to compare the temporal dynamics of object and context visual processing using a saccadic choice task that allows the investigation of an early temporal window of visual processing (120-250 ms). We were particularly interested in the ability of subjects to extract the global information of natural scenes (i.e. is the scene man-made or natural?) and to use the information to guide selective saccades at such latencies. Our results show that human subjects are able to extract this sort of information very quickly, scoring about 73% correct with reliable saccades appearing roughly 160 ms after scene onset. Thus the contextual gist can guide the oculomotor system with a minimal saccadic latency that is somewhat longer than for animal targets (120 ms) but shorter than for vehicle targets (180ms).

The second main aim of the study was to quantify the precise onset of the congruence effect that was reported in an earlier study using a manual go/no-go animal categorization task [25]. In this previous study, even the earliest manual responses were influenced by the nature of the context. Since the saccadic choice task gives access to an earlier temporal window because saccades can be produced faster than manual responses, we were able to precisely time that contextual information start to bias animal detection as early as 160 ms after stimulus onset. This latency fits remarkably well with the first latency at which enough diagnostic scene information is available to allow ultra-rapid context discrimination.

Object processing

The present study replicates the results obtained by Kirchner & Thorpe [31], namely that human subjects can reliably initiate saccades towards the side of an image that contains an animal in just 120-130 ms. The fact that the presence of an animal in an image can be detected very rapidly fits with a number of recent studies showing that the category “animal” is particularly easy to derive on the basis of activity in the temporal lobe. For example, Hung et al [39] demonstrated that object category can be read out rapidly from neuronal responses in monkey inferotemporal cortex, and similarly rapid read-out has recently been demonstrated using intracerebral recordings in human epileptic patients [40]. Clustering techniques have shown that a clear distinction between animate and inanimate objects can be made based either on the pattern of fMRI voxel activation in human cortical areas [42] or on the responses of IT cells in monkeys [41].

This study also extends the results obtained by Crouzet et al. [34] by demonstrating an asymmetry in targeting different object categories. It appears that object

categories are not all equal in this saccadic choice task. Using human face and vehicle categories, Crouzet et al. [34] demonstrated that the very fast saccades towards human faces were not completely under top-down control because they were still biased towards the side where a face is presented even when subjects were instructed to target vehicles. The present study extends this asymmetry to animals and vehicles. The advantage for the animal category is similarly demonstrated by a specific and early pattern of errors when the task requires saccading towards vehicles. A large proportion of early saccades were produced towards the animal distractor suggesting a strong biological saliency for animal objects and a difficulty to control such early saccades.

These results illustrate how the use of the saccadic choice task opens up a behavioural window onto stages of visual processing that were previously inaccessible. For example, VanRullen and Thorpe [10] had shown that when using a manual go/nogo categorization paradigm, subjects can easily switch target category between animals and vehicles. Indeed, performance was remarkably similar in both situations. However, using very similar stimuli, the saccadic choice task used here reveals some very strong asymmetries between animals and vehicles since fast saccades are only observed with animal-targets.

Context processing

The present study also demonstrates that subjects are able to perform this saccadic choice task using man-made or natural environments as targets with a minSRT of 160 ms in both cases. Thus, at such short latencies, there is enough information about scene gist to allow the extraction of coarse scene information, although more detailed scene analysis, such as indoor, outdoor, mountains, sea etc…, might well require more processing [14,43]. Compared to animal processing, the temporal lag of contextual information is surprising given the number of studies that have shown how decisions concerning the nature of the scene can rely on very fast extraction of visual information. For example, Fei-Fei et al. [44] flashed masked gray-scale images using a range of stimulus durations and asked subjects to describe what they had seen. Subjects could report reliably whether the image presented a manmade or a natural environment Even with the shortest presentations, suggesting that this distinction is relatively straightforward for the visual system. This view is reinforced by the fact that relatively simple categorization strategies based on just the global orientation and spatial frequency composition of the image can provide reliable information to discriminate between manmade and natural environments [20,21]. But access to representations depends both

upon stimulus duration and target-scene/object integration time (Vo & Henderson, 2010) and integration time might just be faster for animals than for scene contexts.

Arrival of context information and “congruence” effect

The present data shows a clear influence of context on animal processing, but not on the earliest responses. Indeed, saccades accuracy towards animals became significantly higher when animals were presented in a congruent (vs. incongruent) context, an effect that becomes significant around 160 ms after stimulus onset. It is worth noting that this latency fits perfectly with the minSRT computed with contextual targets that was also seen at 160 ms. On the other hand, this congruence effect was not found with vehicle-targets. Since processing speed for vehicles appears to be slow relatively to animal objects, we could have expected an influence of context from the very first responses towards vehicles. One explanation to the absence of contextual effect could lie in the expertise that we have with vehicles. Most exemplars of the vehicle category are less associated with a specific context than most animals (boats are equally likely to be seen in a port or on the open sea, and cars and planes can also be often seen in purely natural contexts). This differential effect of context among object categories might thus rely on the strength of their level of contextualisation [45]. The processing of an object strongly associated with a specific context will be facilitated when embedded in such appropriate context (a giraffe in the savannah) whereas a conflictual context will generate interference. In contrast, an object that has no strong associations with any specific context will neither take advantage nor suffer from any association. It would be interesting to contrast the animal category with an artefactual category strongly associated with a particular context such as "furniture", that is usually seen indoors.

The results suggest that object and context processing could interact all along the ascending flow of visual processing, even during the first feedforward sweep [24,25,46]. Such interactions would affect the access to object representation when objects are strongly associated with specific contexts. Such observations are inconsistent with the functional isolation model proposed by Henderson and Hollingworth [47,48] where object and context information are proposed to be processed independently without interfering. This finding is also contrasting with the Dynamic Scene Categorization Model in which object will influence scene categorization by modifying global scene statistics [49]. However, it is clear that some object categories (such as faces or animals) appear to be processed even before the extraction of the contextual category. The model of contextual

guidance [50] might well be constrained by an initial processing of global scene properties that would both depends upon stimulus duration and associated integration time [51].

Bar's model [26] postulates that object recognition in the infero-temporal cortex would be influenced by two kinds of top-down modulations based on global, low-spatial-frequency information: a representation of the contextual frame generated by the parahippocampal cortex and a representation of potential candidate objects originating in the prefrontal cortex. The orbitofrontal cortex has been shown to influence the fusiform gyrus at around 130 ms after stimulus onset [30] a latency that is compatible with the temporal course of vehicle detection in our task, but not with animal and face detection that apparently starts earlier. Moreover our study also shows that the contextual frame of a scene is available at the same latency as vehicle objects suggesting parallel processing and access to gist and artefactual object representation at similar latencies, an observation that does not fit easily with a top down influence of the parahippocampal area onto the inferotemporal cortex. More plausible, and based on the parallel processing of the different elements within the visual field [32, 33, 52], object and context information could be processed with a similar time-course (except for specific categories like animals or faces) along the visual path and interact with each other in a bottom-up stream. The physical features they encode could be diagnostic enough to allow the fast decisions observed in our ultra-rapid saccadic task, and their interactions would be the support for the congruency effect. Facilitation would be based on coactivation of selective neuronal responses that have been reinforced through experience, whereas interference would arise from conflictual co-activations [25]. Such learned regularities would not be linked to any conscious strategy. Indeed, it was recently shown that even monkeys trained on object categorization tasks with natural scenes are sensitive to contextual information emphasizing the implicit side of such interactions [53].

An early time dedicated to “special” categories

The fastest saccades towards animals, initiated in the range 130-160 ms, are remarkably unaffected by context congruency. One possibility is that processing of certain key biological stimuli such as animals (and faces) is intrinsically faster than for other stimuli. This result is reminiscent of the bias towards animals that was reported in a change-detection paradigm by New & al. [54]. In that study, subjects were faster and more accurate at detecting changes affecting animals than all other categories of inanimate objects tested by the authors, including vehicles that can also move and be dangerous. For the authors, the data suggest that the human monitoring system was tuned by

ancestral priorities rather than expertise. In other words, monitoring images for the presence of an animal could be somehow "hardwired" [55, 56]. This built-in advantage would lead to the observable latency differences between populations of neurons selective for different object categories. The existence of such latency differences was recently demonstrated in monkey inferotemporal cortex where it was found that neurons responded at latencies that are systematically shorter for primate faces than for other animal faces [57]. This neuronal latency difference could also explain why saccades to animals are observed with a minSRT of 120 ms a value that is longer than the minSRT of 100 ms found for human face targets in the study by Crouzet et al. [34]. One ecological interpretation for these differences would be that animal (or face) detection is such a high priority in everyday life that the visual system has evolved for fast detection of such stimuli.

Conclusion

The present study illustrates how the use of the saccadic choice task opens up a new behavioural window onto early visual processing previously inaccessible. First, it allowed us to reveal a processing time asymmetry between animal and vehicle processing that was invisible with previous manual response protocols. Second, it revealed an earliest availability of pure contextual information, with a latency of 160 ms that fits remarkably to the time at which animal detection started to be biased by context congruency. Overall, these results argue in favor of early bottom-up interactions between object and context processing probably based on fast low-level feature integration.

References :

1. Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. Trends in Cognitive Sciences , 11 (7), 280-289.

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

3. Biederman, I. (1981). On the semantics of a glance at a scene. In: M.K.J.R. Pomerantz (Ed.). Perceptual Organization (pp. 213 -254). Hillsdale: Lawrence Erlbaum.

4. Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. Journal of Cognitive Neuroscience , 15 (4), 600-9.

5. Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review , 113 (4), 766-86.

6. Potter, M., & Faulconer, B. (1975). Time to understand pictures and words. Nature , 253 (5491), 437-8.

7. Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. Nature, 381(6582), 520-522.

8. Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. Vision Res, 40(16), 2187-2200.

9. Rousselet, G. A., Mace, M. J., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. J Vis, 3(6), 440-455.

10. VanRullen, R., & Thorpe, S. J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. Perception, 30(6), 655-668.

11. Mace, M. J., Thorpe, S. J., & Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes: how robust at very low contrasts? Eur J Neurosci, 21(7), 2007-2018.

12. Mace, M., Joubert, O., Nespoulous, J., & Fabre-Thorpe, M. (2009). The Time-Course of Visual Categorizations: You Spot the Animal Faster than the Bird. PloS one , 4 (6).

13. Schmidt, T., & Schmidt, F. (2009). Processing of natural images is feedforward: A simple behavioral test. Attention, Perception & Psychophysics , 71 (3), 594-606.

14. Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. Vision Res, 47(26), 3286-3297.

15. Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the “gist” of real-world natural scenes? Visual Cogn, 12(6), 852-877.

16. Biederman, I. (1995). Visual object recognition. In: S.F.K.D.N. Osherson (Ed.) An invitation to cognitive science (pp. 121-165): MIT Press.

17. Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cognit Psychol, 58(2), 137-176.

18. Greene, M. R., & Oliva, A. (2009). The briefest of glances: the time course of natural scene understanding. Psychol Sci, in press.

19. Torralba, A. (2009). How many pixels make an image? Visual Neuroscience, 26(1), 123-131.

20. Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. Network, 14(3), 391-412.

21. Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 42(3), 145-175.

22. Oliva, A. & Torralba, A. (2006). Building the Gist of a Scene: The Role of Global Image Features in Recognition. Progress in Brain Research: Visual perception, 155, 23-36.

23. Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. Psychol Sci, 5(4), 195-200.

24. Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. Psychol Sci, 15(8), 559-564.

25. Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. J Vis, 8(13), 11-18.

26. Bar, M. (2004). Visual objects in context. Nature Reviews Neuroscience, 5(8), 617-629. doi: 10.1038/nrn1476.

27. Fenske, M., Aminoff, E., Gronau, N., & Bar, M. (2006). Top-down facilitation of visual object recognition: object-based and context-based contributions. Progress in brain research , 155, 3-21.

28. Kirchner, H., Barbeau, E.J., Thorpe,S.J., Régis, J. and Liégeois-Chauvel, C. (2009). Ultra-Rapid Sensory Responses in the Human Frontal Eye Field Region. J Neuroscience, 29(23):7599 –7606.

29. Liu, H., Agam, Y., Madsen, J. R. & Kreiman, G. (2009) Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. Neuron 62, 281-290.

30. Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmidt, A.M., Dale, A.M., Hamalainen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R. and Halgren, E. (2006). Top- down facilitation of visual recognition. Proceedings of the National Academy of Science, 103, 449–454.

31. Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. Vision Res, 46(11), 1762-1776.

32. Rousselet, G., Fabre-Thorpe, M., & Thorpe, S. (2002). Parallel processing in high-level categorization of natural images. Nature Neuroscience , 5 (7), 629-30.

33. Rousselet, G.A., Thorpe, S.J. & Fabre-Thorpe, M. (2004) How parallel is visual processing in the ventral pathway? Trends in Cognitive Sciences, 28(8), 363-70.

34. Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. Journal of Vision, 10(4):16, 1–17, http://journalofvision.org/10/4/16/, doi:10.1167/10.4.16.

35. Bacon-Mace, N., Kirchner, H., Fabre-Thorpe, M., & Thorpe, S. J. (2007). Effects of task requirements on rapid natural scene processing: From common sensory encoding to distinct decisional mechanisms. J Exp Psychol Hum Percept Perform, 33(5), 1013-1026.

36. Brainard, D. H. (1997). The Psychophysics Toolbox. Spat Vis, 10(4), 433-436.

37. Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis, 10(4), 437-442.

38. Smeets, J., & Hooge, I. (2003). Nature of variability in saccades. Journal of Neurophysiology , 90 (1), 12-20.

39. Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science, 310, 863-866.

40. Liu,H., Agam, Y., Madsen,J., & Kreiman, G. Timing, Timing, Timing: Fast Decoding of Object Information from Intracranial Field Potentials in Human Visual Cortex. Neuron, 62(2), 281-290

41. Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. Journal of Neurophysiology , 97 (6), 4296-309.

42. Kriegeskorte,N., Mur,M., Ruff,D., Kiani,R., Bodurka,J., Esteky,H., Tanaka,K., & Bandettini.,P. Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey Neuron, 60(6), 1126-1141

43. Loschky, Lester C. and Larson, Adam M. (2010) The natural/man-made distinction is made before basic- level distinctions in scene gist processing. Visual Cognition, 18(4), 513 – 536

44. Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? J Vis, 7(1), 10.

45. Bar, M., Aminoff, E., Schacter, D.L. (2008). Scenes unseen: the parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se. J Neurosci 28, 8539-8544.

46. Davenport, J. L. (2007). Consistency effects between objects in scenes. Memory & Cognition, 35, 393–401.

47. Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? Journal of Experimental Psychology: General, 127, 398–415.

48. Henderson, J., & Hollingworth, A. (1999). High-level scene perception. Annual Review of Psychology, 50(1), 243–271.

49. Mack, M. L., & Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. Journal of Vision, 10(3):11, 1–11, http://journalofvision.org/10/3/11/, doi:10.1167/10.3.11.

50. Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. Psychological Review, 113, 766–786.

51. Võ, M. L.-H., & Henderson, J. M. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. Journal of Vision, 10(3):14, 1–13, http://journalofvision.org/10/3/14/, doi:10.1167/10.3.14.

52. Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature, 460, 94-97

53. Fize, D., Cauchoix, M., Fabre-Thorpe, M. (in préparation). Large Overlap of Abstract Visual Representations in Human and Monkey

54. New, J., Cosmides, L., & Tooby J. (2007) Category-specific attention for animals reflects ancestral priorities, not expertise. PNAS, 104(42), 16598-603.

55. VanRullen, R. (2007). The power of the feed-forward sweep. Advances in Cognitive Psychology, 3(1-2), 167-176.

56. VanRullen, R. (2009). Binding hardwired versus on-demand feature conjunctions. Visual Cognition, 17(1), 103–119. Psychology Press. doi: 10.1080/13506280802196451.

57. Kiani, R., Esteky, H., & Tanaka, K. (2005). Differences in Onset Latency of Macaque Inferotemporal Neural Responses to Primate and Non-Primate Faces. Journal of Neurophysiology, 94, 1587-1596. doi: 10.1152/jn.00540.2004.

3.1.3 Résumé des principaux résultats

Deux résultats principaux ressortent de cette étude. D'abord, cette étude s'inscrivait dans une thématique de notre équipe de recherche qui cherche à comprendre les interactions entre les traitements de l'objet et le contexte lors de la perception de scènes naturelles. Plus particulièrement, nous cherchions à savoir à partir de quelle latence les informations contextuelles pouvaient influencer la reconnaissance d'objets (si elles le faisaient). Nous avons démontré que cette influence commençait 160 ms après l'affichage des stimuli, une valeur qui correspondait parfaitement à l'arrivée des premières informations permettant de reconnaître l'information contextuelle de la scène. Mais une autre information majeure qui ressort de ce résultat est que durant une fenêtre temporelle précoce, entre 120 et 160 ms, le traitement de l'animal se fait de façon totalement indépendante du contexte.

3.2 Niveaux de catégorisation

Les travaux effectués dans les années 60-70 sur les catégories ont montré que les catégories s'organisent à la fois sur un plan horizontal et sur un plan vertical. Le plan horizontal correspond aux différentes catégories (animaux, véhicules, meubles, etc) et le plan vertical aux différents niveaux de catégorisation (véhicule, voiture, Ford Fiesta). De nombreuses expériences ont montré que les différents niveaux de catégorisation ne sont pas tous équivalents et que le *niveau de base* serait la porte d'entrée privilégiée pour la reconnaissance d'un objet (Mervis et Rosch, 1981). L'accès aux autres niveaux se ferait ainsi à partir du niveau d'entrée. Cette idée forte provient notamment des études de Rosch, où les sujets étaient plus rapides pour catégoriser les objets au niveau de base (oiseau, chien, voiture) qu'aux niveaux superordonnés (véhicule, animal) ou subordonnés (labrador, Ferrari). Selon cette hypothèse, l'accès au niveau superordonné résulterait d'une généralisation (ou abstraction) à partir du niveau de base et l'accès au niveau subordonné résulterait d'une discrimination plus fine utilisant des informations visuelles supplémentaires. Cette architecture sur le plan vertical est toutefois remise en cause dans le cas d'objets atypiques (les manchots parmi les oiseaux par exemple) qui sont catégorisés à la même vitesse au niveau de base et au niveau subordonné (Jolicoeur *et al.*, 1984), le niveau d'entrée pourrait donc parfois se trouver au niveau subordonné.

Nous avons vu dans la section 1.4.1 que les sujets peuvent être à la fois précis et très rapides pour effectuer une tâche de catégorisation de type animal/non-animal. Cette rapidité, en plus de poser certains problèmes aux modèles classiques de reconnaissance, vient directement questionner les études sur les niveaux de catégorisation présentés précédemment. En effet, la catégorie *animal* étant par essence une catégorie superordonnée, il se pourrait donc que les réponses puissent être encore plus rapides pour une

tâche de catégorisation effectuée au niveau de base (par exemple, oiseau/non oiseau ou chien/non chien). Une étude récente montre que ce n'est pas le cas (Macé *et al.*, 2009). En utilisant un protocole go/no-go classique, les auteurs ont comparé différentes conditions manipulant directement le niveau de catégorisation requis par la tâche. Ils ont ainsi montré qu'une catégorisation animal/non-animal pouvait être réalisée 40-65 ms plus rapidement qu'une catégorisation chien/non-chien ou oiseau/non-oiseau. Dans ces deux derniers cas, les non-chiens et non-oiseaux étaient des images contenant d'autres animaux, ceci afin de s'assurer que le niveau de catégorisation requis pour la tâche était vraiment le niveau basique. Cet avantage temporel remet donc largement en question la théorie d'accès aux catégories de Rosch, en démontrant que dans une tâche purement visuelle, la catégorie superordonnée pourrait être le niveau d'entrée.

Cependant, comme il a été montré récemment en comparant les niveaux basiques et subordonnés, des temps de réaction plus rapides ne signifient par forcément que les traitements de l'un soient réalisés avant ceux de l'autres (Mack *et al.*, 2009). Cependant, l'étude de Macé et al. montre bien que les toutes premières réponses pour la catégorie superordonnée apparaissent avant celle pour la catégorie basique. Le protocole de choix saccadique pourrait s'avérer extrêmement intéressant dans ce cadre. Nous savons déjà que la catégorisation animal/non-animal peut tout à fait se faire avec des temps de réaction extrêmement rapides (Kirchner et Thorpe, 2006; Crouzet *et al.*, 2010). En allant observer l'état de l'information très précocement, il se peut que l'information sur la catégorie basique ne soit tout simplement pas encore présente à ce moment.

3.2.1 Résumé de l'étude

Un grand nombre d'études supposent donc que le système visuel accède à la catégorie basique plus rapidement qu'à la catégorie superordonnée ou subordonnée. Cependant, cet avantage a été remis en cause récemment dans le cas de la reconnaissance d'objets (voir notamment : Macé *et al.*, 2009). Nous nous sommes donc intéressés à cette question en utilisant une tâche de choix saccadique. Ce protocole nous permet de comparer les dynamiques temporelles de la reconnaissance d'objets au niveau basique et superordonné, à partir de réponses comportementales très rapides. Quatre conditions ont été comparées dans l'Expérience 1.

Deux conditions qui requièrent un traitement superordonné :

- animal vs. non-animal
- chien vs. non-animal

Deux conditions qui requièrent un traitement basique :

- chien vs. autre-animal
- chien vs. oiseau

Nos résultats montrent que les participants répondent dans une fenêtre temporelle similaire que la tâche se fasse au niveau basique ou superordonné (première réponse vers 120 ms), mais que le niveau de performance, tout à fait correct en discrimination superordonnée (plus de 80%) chute au niveau chance pour le niveau basique. En plus du niveau de catégorisation, il semble important de considérer aussi l'effet de l'homogénéité des exemplaires au sein d'une catégorie, ainsi que la similarité morphologique entre les catégories (deux paramètres importants dans les résultats de recherche visuelle classique, Duncan et Humphreys, 1989). Ceci peut rendre compte du léger avantage observé pour la tâche chien vs. non-animal sur animal vs. non-animal (plus grande homogénéité de la catégorie cible), ainsi que l'avantage de chien vs. oiseau sur chien vs. autre-animal (plus grand homogénéité de la catégorie utilisée en distracteur).

Afin de mieux comprendre cet aspect au niveau basique, l'Expérience 2 a cette fois comparé trois catégories (chien, oiseau, chat) entre elles et a montré que la catégorie oiseau était clairement désavantagée en comparaison aux deux autres. Lorsque les catégories chien et chat sont mises en compétition les sujets sont au niveau chance (ce que l'on pouvait attendre du niveau basique vu les résultats de l'Expérience 1). Au contraire, lorsque les oiseaux sont utilisés en cible, les sujets sont au niveau chance, alors que lorsqu'ils sont utilisés en distracteur, la tâche devient plus facile (autour de 60% de réponse correctes). On a donc un léger biais en faveur des catégories chien et chat comparé à la catégorie oiseau. Ce biais ne pouvant pas s'expliquer par une différence de saillance bas-niveau (en tout cas au sens d'Itti et Koch, Itti et Koch, 2001; Itti *et al.*, 1998), il semble donc que les images de chiens et de chats posséderaient des caractéristiques plus à même d'attirer les saccades rapides (par exemple : quatre pattes, morphologie plus prototypique de la catégorie animal).

Ces résultats, pris dans leur ensemble, supportent donc un modèle de perception rapide dit *coarse-to-fine*, avec un accès rapide à des informations générales permettant de classer l'objet dans sa catégorie superordonnée, puis une progression vers une analyse plus fine de la scène.

3.2.2 Article 4 : At 120 ms you know where the animal is but you don't yet know it's a dog.

At 120 ms you know where the animal is but you don’t yet know it’s a dog

(en préparation)

Chien-Te Wu[1,2]
Sébastien Crouzet[1,2]
Simon J. Thorpe[1,2] & Michele Fabre-Thorpe[1,2]

[1] Université de Toulouse ; UPS ; Centre de Recherche Cerveau et Cognition ; France
[2] CNRS, CerCo ; Toulouse, France

Corresponding Author:
Michele Fabre-Thorpe
Centre de Recherche Cerveau et Cognition
Faculté de Médecine Rangueil,
31062 Toulouse Cedex (FRANCE)

Tel: +5 62 17 28 07
Fax: +5 62 17 28 09
email: michele.fabre-thorpe@cerco.ups-tlse.fr

Article 4 : At 120 ms you know where the animal is but you don’t yet know it’s a dog

Abstract

Earlier studies suggested that the visual system processes information at the basic level faster than at the subordinate or superordinate levels. However, the advantage of the basic category over the superordinate category in object recognition has been challenged recently, and the hierarchical nature of visual categorization is now a matter of debate. In the current study, we addressed this issue using a forced-choice saccadic task in which two images were displayed simultaneously on each trial and participants had to saccade as fast as possible towards the image containing the designated target category. This protocol allows us to compare the temporal dynamics of visual recognition at the basic and superordinate level at very short delays after stimulus onset. Our results revealed that with very short processing times (~120ms), participants were able to perform the task at the superordinate level, but were much worse when the task involved categorization at the basic level. When categorizing an object among different basic levels, saccades started at the same mean latency but accuracy was low, possibly because of the degree of morphological similarity between targets and non-targets. Follow-up computational modeling further confirmed that these behavioral results cannot be predicted by pure bottom-up saliency differences between the images used. Therefore, our results support a coarse-to-fine model of visual recognition. The visual system first gains access to relatively coarse visual representations which provide information at the superordinate level of an object, but additional visual analysis is required to allow more detailed categorization at the basic-level.

Introduction

What can we perceive with just a glance at a scene? How long does it take to recognize an object? These questions have been important topics in studies of human vision since the 70s. Early influential works by Biederman and Potter suggested that human visual system can extract most information from scenes with presentation duration as short as ~ 100 ms (Biederman, Rabinowitz et al. 1974; Potter 1976). Following these findings, numerous studies have further showed that human beings are fast and accurate at categorizing or detecting briefly presented object images that contains animals, vehicles, food objects, human or animal faces among distractors even though these images were just flashed for 20~25ms (Thorpe, Fize et al. 1996; Delorme, Richard et al. 2000; VanRullen and Thorpe 2001; Rousselet, Mace et al. 2003). In these tasks, the earliest selective behavioral responses were observed at 250-270 ms after stimulus onset, a latency range presumably reflecting both the visual processing of images and the production of motor response. In addition to these behavioral evidence, neurophysiological measures also revealed an early scalp potential differentiation correlated with correct visual categorization responses at ~150 ms after image onset (Thorpe, Fize et al. 1996), a latency that can not be shortened further even with extensive training with the stimuli (Fabre-Thorpe, Delorme et al. 2001).

Recently, several studies have further revealed that such rapid visual categorization process can be performed in parallel with at least two images and showed no significant behavioral cost (Rousselet, Fabre-Thorpe et al. 2002; Rousselet, Thorpe et al. 2004). The characteristic of parallel processing has led to the development of a new forced-choice saccade task which aimed to probe the earliest visual decision making process with a much stronger temporal constraints (Kirchner and Thorpe 2006) than traditional paradigm designs (Thorpe,

Fize et al. 1996; Delorme, Richard et al. 2000; VanRullen and Thorpe 2001; Rousselet, Mace et al. 2003). Typically in this task, participants are presented simultaneously with two images in the left and right hemifields, and they are required to make a quick saccade toward the image that contains the pre-specified target object. In general, human beings performed quite well (> 80% correct response) with the earliest reliable saccade to occur at ~120ms post-stimulus when targets are animals and ~100ms post-stimulus when targets are faces (Crouzet, Kirchner et al. in press). Further analysis suggests that this rapid saccadic response toward target images were unlikely to be driven by certain low-level statistical property of the images (Kirchner & Thorpe, 2006). This paradigm therefore serves as a valuable tool to probe the early temporal characteristics of the progressive object representation in the visual system.

In the present study, we aimed to take the temporal advantage of this forced-choice saccade task to address a controversial question on the hierarchy of object categorization (Rosch, Mervis et al. 1976): whether it is easier to access to the basic level (e.g., dogs) information than its corresponding superordinate level (e.g., animals) in the early temporal window of visual processing. A number of earlier studies suggest that basic level representations can be accessed faster than either the superordinate or the subordinate level (Rosch, Mervis et al. 1976; Jolicoeur, Gluck et al. 1984; Murphy and Brownell 1985; Murphy and Wisniewski 1989). This advantage might reflect the fact that it represents the most appropriate abstraction due both to the specificity of its exemplars and their distinctiveness from other objects of the same superordinate category (Murphy and Brownell 1985). A more recent study even suggested that object categorization at the basic level does not require any more processing time than simple object detection (Grill-Spector and Kanwisher 2005; but see Mack, Gauthier et al. 2008). However, several recent findings have challenged the general basic level advantage in visual recognition processes by showing that under certain

circumstances, the processing at the superordinate or subordinate level is faster than at the corresponding basic level (Tanaka and Taylor 1991; Large, Kiss et al. 2004; Joubert, Rousselet et al. 2007; Rogers and Patterson 2007; Mace, Joubert et al. 2009). A potential explanation of these discrepancies could be that most of the earlier studies showing a basic-level advantage use experimental paradigms that involve semantic processing (e.g., verbal report or naming association). For example, subjects may be simultaneously presented with a drawing of an animal together with a word (e.g., "bird", or "animal" or "robin") and be required to determine whether the drawing and word agree. However, in the studies that showed no basic level advantage, subjects typically have to respond whether a particular image belongs to a predefined category based on visual analysis with little involvement of semantic processing (i.e., no object-name association were required). The reduction of demand for semantic processing indeed reverse the basic level advantage of manual response time in visual categorization (Large, Kiss et al. 2004; Mace, Joubert et al. 2009). One caveat for these studies is that manual response times is relatively long (usually lie around a range of 400~500 ms), which makes it difficult to speak about the nature of categorization hierarchies in the early temporal window of visual processing.

In the current study, we aim to further investigate the hierarchical nature of rapid visual categorization using a forced-choice saccade task developed by Kirchner and Thorpe (2006). A key characteristic of this forced-choice saccade tasks is that the saccades made by the subjects to the side containing the target are extremely fast (starting as early as 120 ms after stimulus onset) when compared to traditional manual responses (Guyonneau, Kirchner et al. 2006; Kirchner and Thorpe 2006; Bacon-Mace, Kirchner et al. 2007). Thus it allows us to probe the efficiency of human visual system on object categorization processes within an early temporal window after stimulus onset. To closely simulate the perceptual challenge faced in

real life and reduce potential confounding effects due to stimulus repetition, we used a large pool of natural photographs containing objects which can be either categorized at the superordinate level (e.g., animals) or at the basic level (e.g., dogs, birds, cats). In addition to the traditional comparison between the superordinate vs. basic level categorization, we further investigate the influence of morphological similarity over the processing efficiency of visual categorization. Our results again challenge the traditional hierarchical view which posits that object representation is first accessed at the basic level (or "entry level"). Instead, our results support a coarse-to-fine model of visual recognition: the visual system first gains access to relatively coarse visual representations which provide information at the superordinate level of an object, but additional visual analysis is required to allow more detailed categorization at the basic-level.

General Methods

Apparatus and Procedures

All experiments were performed in a dimly lit room and stimuli were presented with a 19" CRT screen with a resolution of 800*600 pixels and a refresh rate of 100Hz (Iiyama Vision Master PRO 454). Participants held their heads on a chin rest to maintain viewing distance at 60 cm. Image centers were 8.5° from the fixation cross, and the image size was 14°*14°. Stimuli display and eye tracker control was done using Matlab and the Psychophysics Toolbox 3 (Brainard 1997; Pelli 1997). Eye movements were monitored with an IView Hi-Speed eye tracker (SensoMotoric Instruments, Berlin, Germany), which uses infrared tracking system with a sampling rate of 240 Hz. Saccade detection was performed off-line using SMI BeGaze Event Detection and the saccade based algorithm introduced by

Jeroen Smeets and Ignace Hooge (Smeets and Hooge 2003). Only the first saccade (if entering one of the 2 images) after image presentation was analyzed. Its onset is considered as the Saccadic Reaction Time (SRT). Before each block, a 13-point calibration procedure was performed.

The experimental procedures, as illustrated in Figure 1a, were the same in both Experiment 1 and 2: participants performed a saccadic choice task where two images were displayed at the same time on each trial. At the beginning of each trial, observers had to keep their eyes fixated on a black fixation cross for a pseudo-random time interval (800-1600 ms), the fixation cross then disappeared for 200 ms before the presentation of the task-related image pair; this gap period allows faster initiation of saccades (Fischer and Weber 1993; Kirchner and Thorpe 2006). Two task-related natural scene images, one as target and the other one as distractor, were displayed on each side of the screen for 400 ms (Crouzet, Kirchner et al. in press). Participants were required to make a saccade as fast and as accurately as possible to the side where the image contains the target object category. The background color was set to a mid-gray level ([128, 128, 128] in RGB space).

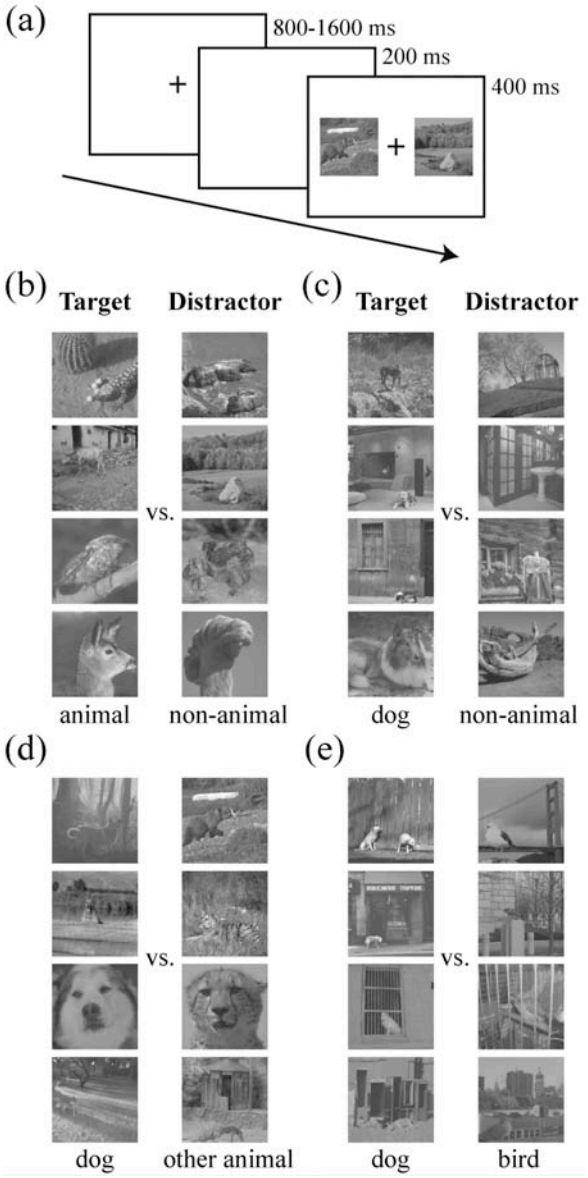

Figure 1 Experimental procedures and sample stimuli in Experiment 1. (a) Each trial started with a fixation cross. After a variable time delay ranging from 800~1600 ms, the fixation cross disappeared for 200 ms and a pair of task-relevant images were presented for 400 ms. Participants would need to make a saccade toward the correct target category as fast and accurate as possible. Here we also showed sample images (left column: target, right column: distractor) used in four different conditions in Experiment 1: (a) Animal vs. non-Animals, (b) Dog vs. non-Animals (c) Dog vs. other Animals, (d) Dog vs. Bird.

Modifier la figure sur deux rangées : a et b et en dessous c, d , e

Data analysis

Minimum saccadic reaction times (Min SRT, Kirchner and Thorpe 2006): To determine the minimum saccadic reaction time for each condition, we first plotted the

histogram of hits and false alarms as a function of saccadic reaction times with a bin size of 10 ms (e.g., a 120 ms bin contains SRT within the range of 115~124 ms). χ^2 tests were calculated for each time bin between hits and false alarms to search for the earliest bin which lead to at least 5 consecutive bins that reached significance level of $p < 0.05$ (i.e., the correct responses exceeded significantly the wrong responses in that saccadic reaction time bin). The first bin in these significant bins was defined as the minimum saccadic reaction time (Min SRT).

Experiment 1

Methods

Participants

Twelve participants (6 males, 6 females, 20-33 year old) performed Experiment 1. All participants provided their informed consents before participating in the experiments.

Stimuli

Experiment 1 aimed at investigating the rapid recognition process at the superordinate level and basic level in the animal category. There were four conditions: animals vs. non-animals (Ani/nonAni), Dog vs. non-animals (Dog/nonAni), Dog vs. other-animals (Dog/oAni) and Dog vs. Bird (Dog/Bird), with the former as the target category and the latter as the distractor category. We considered both Ani/nonAni and Dog/nonAni as conditions involving superordinate level categorization: even if "dog" is considered as a basic category, categorizing dog images vs. non-animal images does not require specific basic level categorization (i.e. subjects might just need to categorize it as an animal to perform the task correctly). On the other hand, the Dog/oAni and Dog/Bird categorization directly involved

object processing at the basic level. A total of 400 images from Corel database and the internet (200 for targets and 200 for distractors) were used for each condition which included 4 runs of 50 trials. The order of conditions were randomized and counterbalanced across participants. Type of background (natural, man-made) and the object size (profile, near, far) were made equal in proportion between every category of images. In addition, we removed the chromatic information from each image and equalized the mean luminance and contrast across images, preventing participants from using certain low-level features (e.g., colors or luminance) as a cue. Figure 1b-1d illustrated some sample images used in Experiment 1.

Results

Accuracy

With short processing times before saccade initiation, results from Experiment 1 provided strong evidence for a categorization advantage at the superordinate level to its corresponding basic level. Although participants in general performed the tasks above chance level (i.e., 50 % of accuracy) for all four conditions (Figure 2), the averaged accuracy for the superordinate categorization (Ani/nonAni & Dog/nonAni collapsed: 80 ± 2%) was significantly higher than that of the basic categorization (Dog/Bird & Dog/oAni collapsed: 60 ± 1%)($t_{11} = 12.96$, $p < 10^{-7}$). This result pattern challenges the common view of a basic level processing advantage.

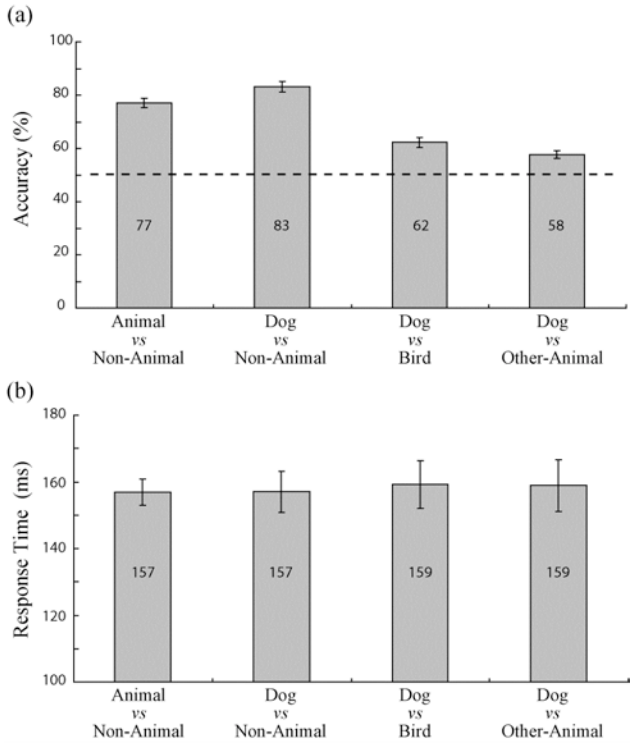

Figure 2 Behavioral result summary in Experiment 1. Results from all four conditions were presented with the top row of the labels referring to target category and the bottom row referring to the distractor category. (a) **Accuracy.** Our results showed that the accuracy was robustly impaired for categorization at the basic level (Dog vs. Bird and Dog vs. Other-Animal) as compared to at the superordinate level (Animal vs. Non-Animal and Dog vs. Non-Animal). (b) **Response time.** However, the mean reaction times did not differ significantly between the two levels of categorization.

Processing Speed

Concerning the speed of processing, unlike accuracy, there was interestingly no significant difference of the mean reaction times between the two level of categorization (Superordinate level: 157 ms vs. basic level: 159 ms, $t_{11} = 0.48$, n.s.). To further evaluate the capability of the visual system to categorize an object within the earliest processing time window, we calculate the minimum reaction time (MinSRT, see Methods) for each condition. The measure of MinSRT reflects the shortest processing time for the brain to differentiate between two categories reliably and initiate the saccade (i.e., correct response significantly exceeds incorrect responses, see General Methods). As illustrated in Figure 3, all conditions,

except the Dog/oAni condition, showed the same MinSRT of 120 ms. These results suggested that when the processing time is as short as 120 ms, the visual information have been processed enough to access a representation of the category at the superordinate level. However, no significant MinSRT was extracted in the Dog/oAni condition, indicating that the categorization performance at the basic level was never stably differed significantly from chance for a significant amount of continuous bins. In other words, the performance fluctuated quite a bit when processing time was kept below 300ms, which is generally the case in the choice saccade task below 300 ms (See Figure 3). To summarize, in contrast to previous findings using manual response task which showed a ~20 ms time delay for basic vs. superordinate categorizations (Mace, Joubert et al. 2009), our results clearly showed that when responses are produced earlier in time, as in the saccadic choice task, a categorization task is still possible at the superordinate level but not at the basic level.

One interesting finding in Experiment 1 was that, unlike Dog/oAni condition, we were able to extract a significant MinSRT from the Dog/Bird condition even though it involves a categorization process at the basic level. One possible explanation was that the homogeneity of morphological difference between dogs (e.g., four legs, fur) and birds (e.g., wings, feathers) reduced the difficulty level of categorization, a well-known effect in computer vision and visual search (Duncan and Humphreys 1989). Whereas in the case of Dog/oAni, the distractors were a variety of different animals, which occasionally shared similar morphological features as dogs (e.g., most mammals) and increased the difficulty levels for the categorization process. A very recent study indeed reported that when a distractor shared more common features with the target object, it tended to increase the possibility for a false-alarm response (Mace, Joubert et al. 2009). Therefore, in Experiment 2, we investigated how

the morphological similarity would modulate categorization performance at the basic level in the forced choice saccadic task.

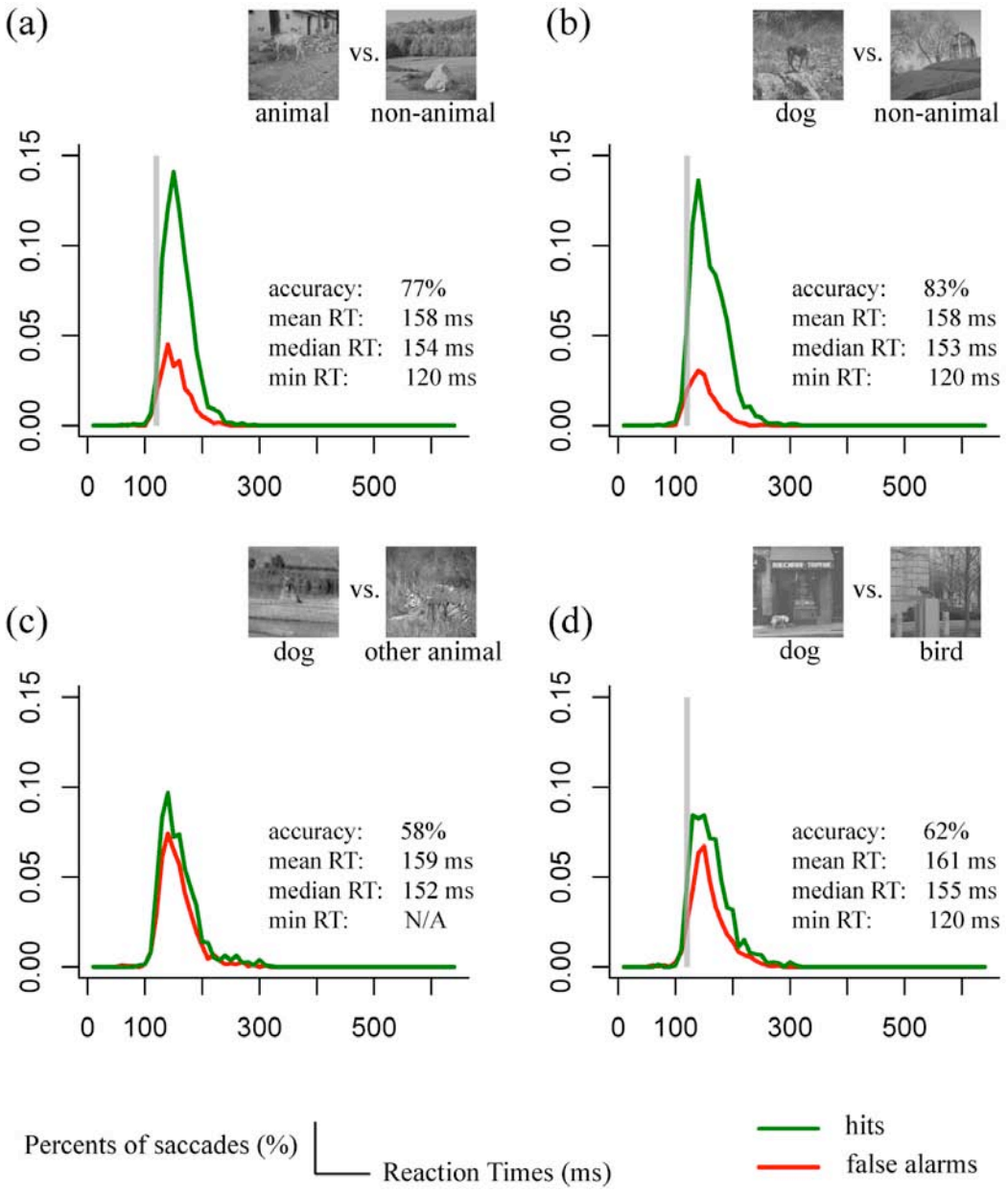

Figure 3 Reaction time distribution and corresponding minimum saccadic reaction times in Experiment 1. Here we plot the histograms of correct responses as a function of reaction time using 20 ms bin size for different conditions in Experiment 1. We were able to extract statistically significant values of minimum reaction times for the condition of Animal vs. Non-Animal (120 ms), Dog vs. Non-animal (120 ms) and Dog vs. Bird (120 ms). However, the number of correct vs. incorrect response were never found to differ significantly for at least 5 continuous bins in the Dog vs. Other-Animal task.

Experiment 2

Methods

Participants

Twelve participants (4 males, 8 females, 21-33 year old) performed Experiment 2. All participants provided their informed consents before participating the experiments.

Stimuli

Experiment 2 aimed at investigating the influence of target/distractor morphological similarity in object categorization at the basic level. There were six conditions: Dog target vs. Bird distractor (Dog/Bird), Bird target vs. Dog distractor (Bird/Dog), Dog target vs. Cat distractor (Dog/Cat), Cat target vs. Dog distractor (Cat/Dog), Cat target vs. Bird distractor (Cat/Bird), Bird target vs. Cat distractor (Bird/Cat). As in Experiment 1, each condition contained 200 trials (400 images, 200 targets and 200 distractors). Contrary to Experiment 1, targets and distractors were paired on each trial according to object size (profile vs. profile, far vs. far, near vs. near). The order of conditions were randomized and counterbalanced across participants. Again, all images were achromatic and equalized for global luminance and contrast.

Results

Accuracy

As hypothesized, there was a strong influence of target/distractor morphological similarity over the basic-level categorization performance across all conditions ($F_{5,66} = 28.3$, $p < 10^{-6}$). Performance was significantly better with dog or cat targets among bird distractors

than when they were categorized against each other (Dog/Bird vs. Dog/Cat, $t_{11} = 7.06$, $p < 0.0001$; Cat/Bird vs. Cat/Dog, $t_{11} = 7.40$, $p < 0.0001$). Interestingly, however, when birds were targets and distractors were dogs or cats, the performance dropped to chance level (Bird/Cat: 47% ± 2%, Bird/Dog: 47% ± 2%, Figure 4a). In other words, there seemed to be a perceptual saliency bias toward dogs or cats as compared to birds, that counteracts with influence of morphological differences between them.

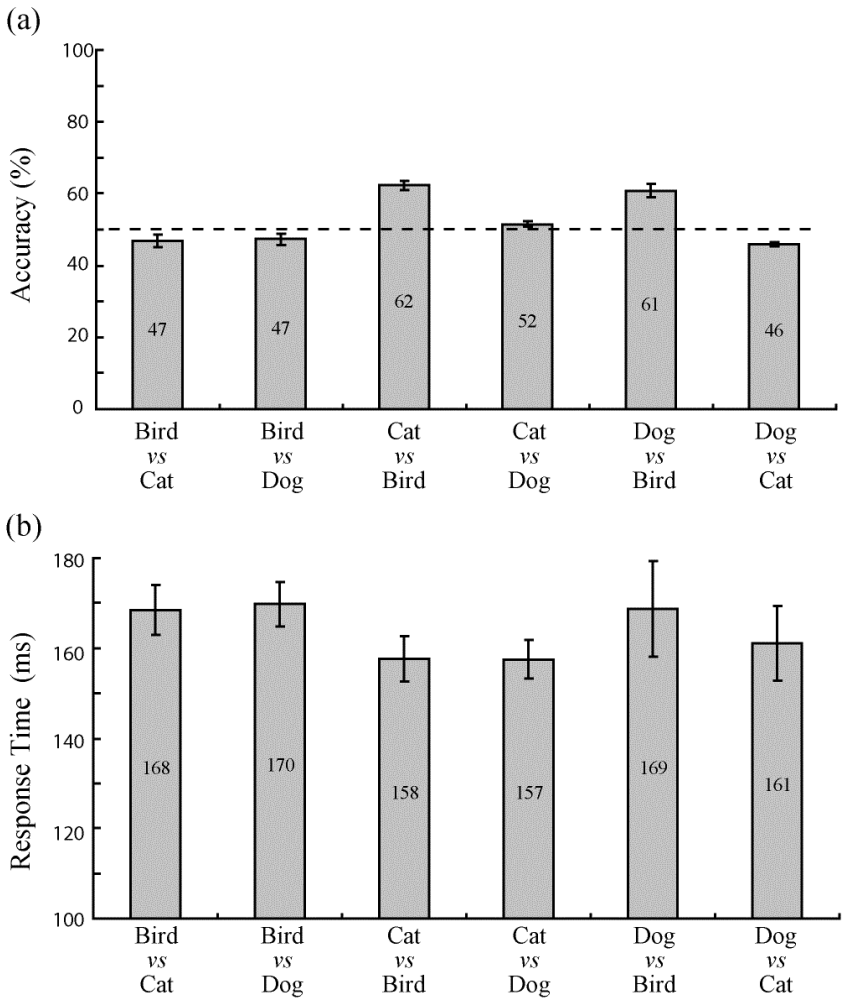

Figure 4 Summary of behavioral result in Experiment 2. (a) **Accuracy.** Performance was above chance level when either Dogs or Cats were targets among bird distractors, but not the other way round. When categorizing between dogs and cats, performance was also at chance level. (b) **Response time.** Again, the mean reaction times did not differ significantly across different conditions.

Processing Speed

Concerning the speed of processing, we found no influence of morphological similarity on different basic level categorization, as indicated by the absence of significant difference of the mean reaction times across conditions ($F_{5,66}$ = 0.54, n.s.). Furthermore, Dog/Bird (MinSRT: 110 ms) and Cat/Bird (MinSRT: 120 ms) were the only two conditions in which we could obtain statistically significant MinSRT (Figure 5). This result pattern, together with the accuracy data, suggests that when processing time is robustly reduced by using saccadic responses rather than manual responses, it is only possible to categorize objects at the basic level above chance level when the morphological difference between the target and distractor categories exceeds a certain threshold. However, in line with the observed accuracy data, we did not obtain a statistically significant MinSRT when birds were targets among either dogs or cats (Bird/Dog and Bird/Cat conditions). Moreover, in both conditions, the false-alarm rates slightly exceed the hit rates across all recorded saccade RT latency (Figure 5e & 5f).

The chance level performance in both the Bird/Dog and Bird/Cat conditions was a bit unexpected considering that they share the same degree of morphological similarity as the Dog/Bird and Cat/Bird conditions, indicating that some additional factors were influencing participants' saccadic behavior. Considering the response latency were very short in the current saccadic task (mean RT in Experiment 2: 158 ~170 ms) as compared to manual responses in previous studies (~500 ms), it is reasonable to suspect that there might be some inherent low-level saliency differences between the selected images of different categories (e.g., cats or dogs were more "salient" than birds) that causes certain perceptual bias toward a specific category, even though they were all transformed into gray-scaled images and equalized for mean luminance and global contrast. Therefore, in the following experiment, we tried to address this issue by comparing saccadic performances of our participants to those of a

computational model based only on the low-level saliency value defined by the configuration of local contrast (Itti and Koch 2000; Itti and Koch 2001).

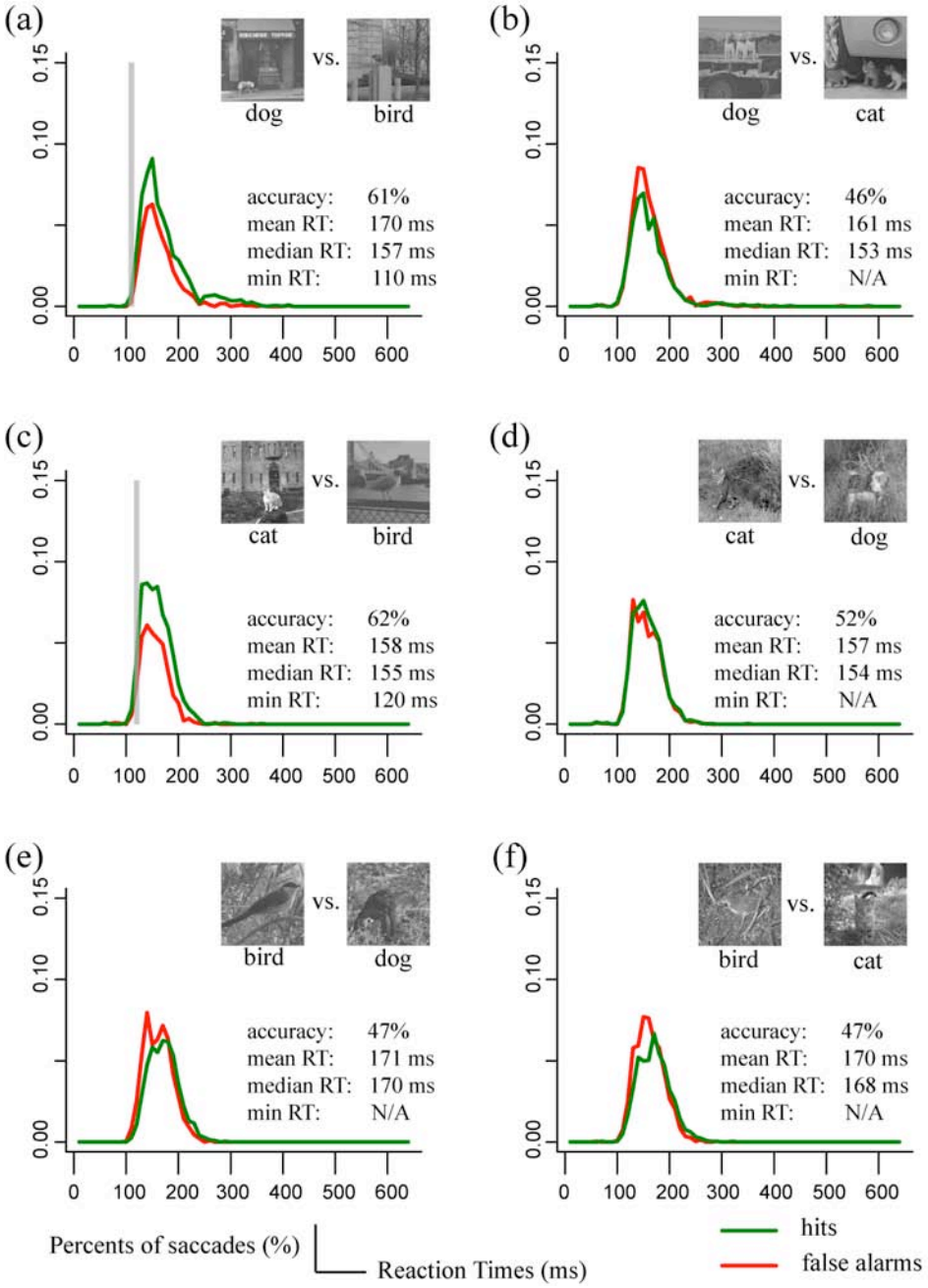

Figure 5 Reaction time distribution and corresponding minimum reaction times in Experiment 2. Histograms of correct responses are plotted as a function of reaction time using 20 ms bin size for the different conditions of Experiment 2. We were able to extract statistically significant measures of minimum reaction times only for two conditions: condition (a) with Dog as targets vs. Bird (minSRT=110 ms) and condition (b) with Cat as targets vs. Bird (minSRT=120 ms).

Saliency simulation

Since the latencies of saccadic reaction times in the current study were very short (mean SRT: 158~171 ms; median SRT: 152~170 ms), one might argue that some inherent differences of bottom-up saliency values between images may play a role in triggering saccades towards a given set of images. In other words, participants' saccades would be automatically attracted toward the image with a higher low-level saliency value. In order to rule out the potential confounding influence of low-level salient features of images on the observed results, we used the Saliency Toolbox 2.2 (Walther and Koch 2006) running under Matlab to simulate the saccadic behavior under the pure influence of low-level saliency values.

Methods

Using Saliency Toolbox 2.2, the input image was processed for low-level feature representation (pixel intensity, orientations and color) at multiple scales. Because all images used in the current study were in gray-scale, we calculated only the intensity and orientation features. The resulting feature maps were integrated to form a saliency map which predicted the locations for the first saccade based on the highest saliency value.

We used the algorithm to simulate each trial that participants experienced during the real experiment. For each trial, the input image was almost the same as what participants observed during the real experiment (2 images on a gray background), with only one exception that the borders of images were smoothed to prevent any inherent bias towards the borders. The location of the first saccade based on the highest saliency value was recorded. The algorithm has no knowledge of the task requirement and thus would make the saccades simply

depending on the most salient location among the two images. But according to the task at hand, if the location was on the target side then it would be taken as a correct response, otherwise as an incorrect response. Such a simulation could provide insights about the potential contribution of low-level features to the triggering of the saccadic responses observed in the current study.

Results

Ideal observer analysis based on pure bottom-up saliency

Results from the current simulation revealed no significant inherent low-level saliency bias in favor of a specific category in either Experiment 1 or 2 (Table 1). In Experiment 1, the performance of the salience toolbox ranged from 47~55%, indicating that local saliency value defined by configuration of local contrasts provided no valid cues about the location of the target images. In other words, there were no significant biases of the low level salience in favor of either the target or the distractor category in the image sets in Experiment 1. In Experiment 2, however, there seemed to be a slight tendency for the Cat category to be more salient than either the Dog (65%) or Bird (62%) category. Although this might explain the positive result for the Cat/Bird condition (model performance: 62% vs. human performance: 62%) and the null result for the Bird/Cat condition (human performance: 47%), there is still a huge discrepancy between the performance of the salience model and the human performance in the Cat/Dog condition (model performance: 65% vs. human performance: 52%). Low level saliency cannot explain chance level when cats were targets among dogs.

How well the model predicts participants' choices?

A fair measure to quantify the potential influence of the low-level salience values of the images over participants' saccadic performances would be to analyze if the saliency model can predict participants' performances in our two experiments on a trial by trial analysis. In other words, for each trial we could determine whether the participant saccadic direction agrees with the saccadic direction of the salience model, independently of the correctness of the response. In the most extreme case, if participants were only using low-level salience features as cues to orient their saccades, the predictability of the model should be perfect (i.e., 100%). As listed in Table 2, the predictability of the model over participants' saccade direction ranged from 50~54% which was nearly chance level. It might still be possible that the low-level saliency values attract saccades more strongly in the early response time window than in the late response time window. If this were the case, the predictability of the model would decrease as the response time increases. In the current study, however, the performance predictability of the model did not show any significant difference across increasing response times. Therefore, these results altogether suggest that the influence of low level salience values, if any, cannot account for the observed performance of participants and thus that local saliency stands as a poor descriptor of the present behavioral results.

Table 1 Salience simulation performance for Exp 1 & Exp 2

Experiment 1			
Ani/nonAni	Dog/nonAni	Dog/oAni	Dog/Bird
52%	55%	47%	47%

Experiment 2					
Dog/Bird	Dog/Cat	Cat/Bird	Cat/Dog	Bird/Dog	Bird/Cat
48%	37%	62%	65%	53%	38%

Table 2 Salience prediction on Participants' performance for Exp 1 & Exp2

Experiment 1			
Ani/nonAni	Dog/nonAni	Dog/oAni	Dog/Bird
54%	55%	50%	53%

Experiment 2					
Dog/Bird	Dog/Cat	Cat/Bird	Cat/Dog	Bird/Dog	Bird/Cat
50%	50%	54%	53%	53%	53%

Discussion

The aim of current study was to investigate the dynamic of object representation within an early time window of rapid object recognition process. A recent study using go/no-go paradigm with manual response showed that, contrary to the common view that the basic level information is accessed first in object representation (Rosch, Mervis et al. 1976), it is faster to categorize the object as an animal (superordinate) than as a dog (Mace, Joubert et al. 2009). Using a forced-choice saccade task involving much faster behavioral responses, our study showed that the access to basic level information is very limited in the early temporal window ~ 120 ms, which is not the case for the superordinate level information. Consistent with a previous study using a similar paradigm design (Kirchner and Thorpe 2006), our results showed that participants were very good at making saccades towards the image containing animals among various distractor images containing no animals. When restricting the target category to be dog images, keeping the same distractors, performance was very comparable, with even a slight increase of accuracy. The main change happened when subjects had to saccade towards dog images among distractors containing other animals (basic level categorization). Then the performance dropped close to the chance level. Restricting the distractor stimuli to only one category (for example here birds) was not sufficient to recover a good level of performance, which was still close to chance. These results demonstrate that superordinate level information is available before the information concerning basic level

category. In other words, that there is a time when you can "spot" the animal, but then need some additional processing to be able to determine the species (and probably even more to determine the identity).

A surprising results, investigate further in Experiment 2, was that different categories at the basic level leaded to different level of performance. Participants were better at making saccade towards dogs and cats than at birds. A reasonable initial explanation would be that there exist some low-level features in our image sets that might bias the saccadic responses. However, this hypothesis was ruled out by a simulation based on a model of saliency map (Itti and Koch 2000; Itti and Koch 2001). This simulation demonstrated that none of our results can be explained by a low-level saliency bias based on local contrast, and furthermore, that this model cannot predict participants' response pattern in this task, even for the fastest saccades (i.e., the first quartile of the saccade responses in the SRT distribution) on which low-level saliency value may presumably render the most robust influence.

Distractor effect on accuracy, not on SRT

In a previous study using the same protocol, large differences in processing times have been demonstrated between different target categories, with minimum SRT of 100 ms for faces, 120 ms for animals, and 140 ms for vehicles (Crouzet, Kirchner et al. in press). Here surprisingly, all the different conditions leaded to very similar processing time (120 ms), a value perfectly consistent with previous studies using animal targets and saccadic choice (Guyonneau, Kirchner et al. 2006; Kirchner and Thorpe 2006; Bacon-Mace, Kirchner et al. 2007; Crouzet, Kirchner et al. in press). This observation is extremely interesting because it suggests that the time when subjects initiate a saccade is not related to the difficulty of the task they are performing but rather on the nature of the target category, the difficulty of the task

being manifested only in the accuracy level. It suggests that the saccade is initiated when the evidences for one alternative is reached independently of the other alternative, the distractor can thus only cause a false alarm (this FA rate being higher when distractors are more similar to the targets), but will not have any influence on the delay for the target to reach its decision threshold. In the decision-making framework, it suggests that in the saccadic choice task, the competition between alternatives is driven by a process of independent accumulation of information rather than mutual inhibition between alternatives, which is often assumed in most of protocols (Smith and Ratcliff 2004).

Neuronal latencies and the saccadic choice task

Furthermore, the fact that there was no difference of RT between conditions here is perfectly consistent with the idea that RTs are mainly driven by the latency of neurons selective for the target objet category (Crouzet, Kirchner et al. in press). Here, because the object category is always an animal, RTs do not vary between conditions. If we consider (i) that the decision in the saccadic choice task is based on the read-out of sensory populations selective for each objet category involved (Glimcher 2003; Gold and Shadlen 2007; Heekeren, Marrett et al. 2008) (as it is often assumed, Glimcher, 2003; Gold & Shadlen, 2007; Heekeren et al., 2008) (ii) that this latency differs between object categories (Kiani, Esteky et al. 2005) (for example in IT, Kiani et al., 2005), then a possible underlying mechanism to explain the present results could be that populations selective for different basic categories are less separable than for superordinate ones (as it seems suggested by cluster analysis, Kiani, Esteky et al. 2007; Kriegeskorte, Mur et al. 2008) as it seems suggested by cluster analysis. The performance in the choice saccade task would thus manifest the acuity of the oculomotor system to select pertinent sensory population to listen.

From superordinate to basic category

Evidence that subordinate recognition takes more time and involve additional processing than basic has been demonstrated several times (Gauthier, Anderson et al. 1997). Because basic level is considered as the entry level for recognition, superordinate recognition is also supposed to take more time (some ref of the intro). However, a coarse-to-fine process could also be relevant for visual processing. Following from this theory, visual analysis at the superordinate level would not be an abstraction from the basic recognition but rather its prerequisite. In a recent study, Fei-Fei et al. presented natural images for various amount of time (followed by a mask) and asked subjects to write what they saw. The analysis of their descriptions as a function of “integration” time revealed that the recognition at the basic level needed more time than at the superordinate level (Fei-Fei, Iyer et al. 2007). A recent TMS study has directly tackled with the underlying brain mechanisms of basic level recognition, showing that it was significantly impaired when pulses were applied 100 ms (first feedforward sweep) but also 220 ms (recurrent processing) after stimulus onset (Camprodon, Zohary et al. 2010). All together, these results strongly suggest that an animal can be recognize at its superordinate level in a single feedforward sweep (which would be the window offered by the saccadic choice task, and then additional recurrent processing would be needed for any more precise analysis, like basic and subordinate level.

Reference

Bacon-Mace, N., H. Kirchner, et al. (2007). "Effects of task requirements on rapid natural scene processing: From common sensory encoding to distinct decisional mechanisms." J Exp Psychol Hum Percept Perform **33**(5): 1013-1026.

Biederman, I., J. C. Rabinowitz, et al. (1974). "On the information extracted from a glance at a scene." J Exp Psychol **103**(3): 597-600.

Brainard, D. H. (1997). "The Psychophysics Toolbox." Spat Vis **10**(4): 433-436.

Camprodon, J. A., E. Zohary, et al. (2010). "Two phases of V1 activity for visual recognition of natural images." J Cogn Neurosci **22**(6): 1262-1269.

Crouzet, S. M., H. Kirchner, et al. (in press). "Fast saccades toward faces: Face detection in just 100 ms." Journal of Vision.

Delorme, A., G. Richard, et al. (2000). "Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans." Vision Res **40**(16): 2187-2200.

Duncan, J. and G. W. Humphreys (1989). "Visual search and stimulus similarity." Psychol Rev **96**(3): 433-458.

Fabre-Thorpe, M., A. Delorme, et al. (2001). "A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes." J Cogn Neurosci **13**(2): 171-180.

Fei-Fei, L., A. Iyer, et al. (2007). "What do we perceive in a glance of a real-world scene?" J Vis **7**(1): 10.

Fischer, B. and H. Weber (1993). "Express Saccades and Visual Attention." Behav Brain Sci **16**(3): 553-567.

Gauthier, I., A. W. Anderson, et al. (1997). "Levels of categorization in visual recognition studied using functional magnetic resonance imaging." Curr Biol **7**(9): 645-651.

Glimcher, P. W. (2003). "The neurobiology of visual-saccadic decision making." Annu Rev Neurosci **26**: 133-179.

Gold, J. I. and M. N. Shadlen (2007). "The neural basis of decision making." Annu Rev Neurosci **30**: 535-574.

Grill-Spector, K. and N. Kanwisher (2005). "Visual recognition." Psychol Sci **16**(2): 152-160.

Guyonneau, R., H. Kirchner, et al. (2006). "Animals roll around the clock: the rotation invariance of ultrarapid visual processing." J Vis **6**(10): 1008-1017.

Heekeren, H. R., S. Marrett, et al. (2008). "The neural systems that mediate human perceptual decision making." Nat Rev Neurosci **9**(6): 467-479.

Itti, L. and C. Koch (2000). "A saliency-based search mechanism for overt and covert shifts of visual attention." Vision Res **40**(10-12): 1489-1506.

Itti, L. and C. Koch (2001). "Computational modelling of visual attention." Nat Rev Neurosci **2**(3): 194-203.

Jolicoeur, P., M. A. Gluck, et al. (1984). "Pictures and names: making the connection." Cognit Psychol **16**(2): 243-275.

Joubert, O. R., G. A. Rousselet, et al. (2007). "Processing scene context: fast categorization and object interference." Vision Res **47**(26): 3286-3297.

Kiani, R., H. Esteky, et al. (2007). "Object category structure in response patterns of neuronal population in monkey inferior temporal cortex." J Neurophysiol **97**(6): 4296-4309.

Kiani, R., H. Esteky, et al. (2005). "Differences in onset latency of macaque inferotemporal neural responses to primate and non-primate faces." J Neurophysiol **94**(2): 1587-1596.

Kirchner, H. and S. J. Thorpe (2006). "Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited." Vision Res **46**(11): 1762-1776.

Kriegeskorte, N., M. Mur, et al. (2008). "Matching categorical object representations in inferior temporal cortex of man and monkey." Neuron **60**(6): 1126-1141.

Large, M. E., I. Kiss, et al. (2004). "Electrophysiological correlates of object categorization: back to basics." Cogn Brain Res **20**(3): 415-426.

Mace, M. J., O. R. Joubert, et al. (2009). "The time-course of visual categorizations: you spot the animal faster than the bird." PLoS One **4**(6): e5927.

Mack, M. L., I. Gauthier, et al. (2008). "Object detection and basic-level categorization: sometimes you know it is there before you know what it is." Psychon Bull Rev **15**: 28-35.

Murphy, G. L. and H. H. Brownell (1985). "Category differentiation in object recognition: typicality constraints on the basic category advantage." J Exp Psychol Learn Mem Cogn **11**(1): 70-84.

Murphy, G. L. and E. J. Wisniewski (1989). "Categorizing objects in isolation and in scenes: what a superordinate is good for." J Exp Psychol Learn Mem Cogn **15**(4): 572-586.

Pelli, D. G. (1997). "The VideoToolbox software for visual psychophysics: transforming numbers into movies." Spat Vis **10**(4): 437-442.

Potter, M. C. (1976). "Short-term conceptual memory for pictures." J Exp Psychol Hum Learn **2**(5): 509-522.

Rogers, T. T. and K. Patterson (2007). "Object categorization: Reversals and explanations of the basic-level advantage." J Exp Psychol Gen **136**(3): 451-469.

Rosch, E., C. B. Mervis, et al. (1976). "Basic objects in natural categories." Cognit Psychol **8**: 382-439.

Rousselet, G. A., M. Fabre-Thorpe, et al. (2002). "Parallel processing in high-level categorization of natural images." Nat Neurosci **5**(7): 629-630.

Rousselet, G. A., M. J. Mace, et al. (2003). "Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes." J Vis **3**(6): 440-455.

Rousselet, G. A., S. J. Thorpe, et al. (2004). "How parallel is visual processing in the ventral pathway?" Trends Cogn Sci **8**(8): 363-370.

Smeets, J. B. and I. T. Hooge (2003). "Nature of variability in saccades." J Neurophysiol **90**(1): 12-20.

Smith, P. L. and R. Ratcliff (2004). "Psychology and neurobiology of simple decisions." Trends Neurosci **27**(3): 161-168.

Tanaka, J. W. and M. Taylor (1991). "Object categories and expertise: Is the basic level in the eye of the beholder?" Cognit Psychol **23**: 457-482.

Thorpe, S., D. Fize, et al. (1996). "Speed of processing in the human visual system." Nature **381**(6582): 520-522.

VanRullen, R. and S. J. Thorpe (2001). "Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects." Perception **30**(6): 655-668.

Walther, D. and C. Koch (2006). "Modeling attention to salient proto-objects." Neural Netw **19**(9): 1395-1407.

3.2.3 À quoi correspond le temps nécessaire pour accéder à la catégorie basique ?

Il semble donc que le système visuel ait besoin de plus de temps pour accéder à la catégorie basique qu'à la catégorie superordonnée. A quoi correspond donc ce temps additionnel ? Différentes explications sont envisageables à différentes niveaux.

Feedback ?

Une première hypothèse serait le besoin de feedback. De nombreuses études ont suggéré la possibilité de réaliser une catégorisation animal/non-animal à partir d'une seule vague feedforward qui traverserait la voie visuelle ventrale (Bacon-Macé *et al.*, 2005; VanRullen et Thorpe, 2002; Thorpe *et al.*, 1996; VanRullen, 2007). Le temps additionnel nécessaire pour la catégorisation au niveau basique pourrait donc correspondre au besoin de feedback. Cette hypothèse semble être confirmée par une étude récente de TMS. Dans celle-ci, les sujets devaient discriminer des images pour les classer en oiseau ou mammifère. Les impulsions magnétiques de TMS étaient envoyées sur V1 à différentes latences après l'affichage des images (Camprodon *et al.*, 2010). Les résultats montrent que les performances de sujets chutaient significativement à deux moments précis : lorsque les impulsions étaient placées 100 ms et 220 ms après l'affichage des images. La valeur de 100 ms pourrait correspondre à la première vague feedforward, celle de 220 ms à des boucles feedback. Si l'on accepte l'hypothèse que ces latences correspondent vraiment aux latences des vagues feedforward et feedback, ces résultats suggèrent que dans une catégorisation au niveau basique, la perturbation du feedback porte atteinte à la performance. Cependant, afin d'être certain de la nécessité spécifique de feedback pour le niveau basique et pas le niveau superordonné, une étude similaire mais utilisant aussi une catégorisation au niveau superordonné est nécessaire.

Attention ?

Une deuxième hypothèse (qui n'est pas antinomique à la première) serait que le traitement au niveau basique nécessiterait la mise en place de mécanismes attentionnels. Comme nous l'avons vu, la catégorisation animal/non-animal peut être réalisée sans attention (Li *et al.*, 2002; VanRullen *et al.*, 2004). Peut-on aussi accéder à la catégorie basique sans attention ? Ou alors est ce que la mise en place de mécanismes attentionnels correspond au temps additionnel mis en avant par notre étude et celle de Macé et al. ? Des études manipulant le niveau de catégorisation en double tâche s'avèreront très intéressantes pour répondre à ces questions.

Le temps de prise d'information ?

Une étude récente a indirectement contribué à explorer cette question (Fei-Fei *et al.*, 2007). Les auteurs demandaient à des sujets d'écrire ce qu'ils avaient perçu d'une image présentée à l'écran. Ces images étaient présentées pour une certaine durée, et immédiatement suivies par un masque. L'analyse consistait ensuite en une analyse textuelle de leur retranscription en fonction de l'intervalle de temps entre le début de l'affichage et l'apparition du masque. De façon intéressante, des animaux étaient parfois présents dans les images. Les résultats montrent clairement que pour les intervalles courts (environ 100 ms), les sujets étaient capables de dire qu'ils avaient vu un animal, mais que pour dire que celui-ci était un chien ou un oiseau, il leur fallait un intervalle plus long (environ 500 ms). La reconnaissance au niveau basique demanderait donc plus de temps pour récupérer de l'information que la reconnaissance au niveau superordonné.

3.3 Conclusions

Ces deux expériences nous ont d'abord permis d'en savoir plus sur les processus que nous avons étudiés, ceci en testant différents effets (influence du contexte et accès au niveau basique) dans une fenêtre temporelle précoce. Mais ces résultats nous renseignent aussi sur la nature des mécanismes en jeu durant la tâche de choix saccadique. Ainsi, la différence entre catégories déjà observée dans l'article 1 est ici confirmée dans l'article 3 où un biais vers les animaux a été observé lorsqu'ils étaient présentés face à des véhicules. De façon importante, cet avantage pour la catégorie animal était associé à une indépendance aux effets contextuels. Il semble donc que des mécanismes spécifiques soient mis en jeu, permettant un traitement spécifique de certaines catégories (animal dans l'article 3). Ces traitements précoce permettent de les détecter très rapidement, avant même que les informations globales contextuelles n'aient pu entrer en jeu. Ces traitements rapides restent cependant assez brutes et imprécis, car comme le montre les résultats de l'article 4, l'accès à la catégorie basique n'est pas possible à des latences si courtes.

Ainsi, il est important de noter que la tâche de choix saccadique est un outil de choix pour l'étude des processus visuels rapides mais ne peut pas être utilisée pour toutes les questions expérimentales. Par exemple, puisque la catégorie basique n'est pas accessible rapidement, on peut supposer que le traitement de l'identité (ou de l'item unique) ne le sont pas non plus. La tâche de choix saccadique est donc un protocole d'étude des traitements rapides, qui pourraient correspondre aux traitements ne nécessitant qu'une seule vague d'information feedforward. Elle peut donc permettre d'étudier et de mettre en évidence des différences précises entre les traitements pouvant

être réalisés en une seule passe. Par contre, ce protocole sera moins pertinent pour l'étude des différences de temps de traitement dans des tâches plus précises, car il ne permettra pas d'avoir des temps de référence précis du fait du faible niveau de performance des sujets. Ces résultats sont assez logiques si on les replace dans le cadre du rôle des mouvements oculaires dans la vie réelle. Leur but est justement d'amener sur la fovéa les régions du champ visuel potentiellement intéressantes. Si l'analyse visuelle était déjà complète avant d'initier le mouvement oculaire, alors il n'y aurait aucun intérêt à aller chercher plus d'information. Au contraire, la détection rapide d'information potentiellement intéressante permettra d'orienter le regard au plus vite, afin de pouvoir poursuivre une analyse plus poussée. Ces informations potentiellement intéressantes peuvent bien évidemment être un mouvement ou un flash lumineux, mais aussi, et c'est la thèse soutenue dans ce mémoire, des objets d'intérêt.

Chapitre 4

De la perception à l'action : la décision

Sommaire

4.1 Problématique

Dans la section 2.2, nous avons vu que des saccades pouvaient être initiées vers une cible, lorsque celle-ci est un visage, en seulement 100 ms. Il est communément admis que le délai entre le moment du choix et le mouvement effectif des yeux serait d'environ 20 ms. Pour les saccades les plus rapides, l'information de décision doit donc être disponible seulement 80 ms après l'affichage des stimuli. Un mécanisme de décision extrêmement simple doit donc être à l'œuvre ici.

Nous avons vu dans cette même étude que la grande rapidité pour détecter les visages était associée à une grande difficulté à répondre vite et correctement lorsque la cible était un véhicule et le distracteur un visage. Ce que l'on a appelé le biais vers les visages. Une hypothèse simple, s'inspirant des résultats d'une étude d'électrophysiologie récente (Kiani *et al.*, 2005), est que ces réponses rapides et ce biais seraient causés par une différence de latence entre les neurones sensoriels sélectifs aux visages et

ceux sélectifs aux véhicules. Ces différences de latences entre populations de neurones pourraient expliquer à la fois les différences de temps de réaction observées entre les différentes catégories dans la tâche de choix saccadique, ainsi que l'aspect automatique des toutes premières saccades.

Pour tenter de valider cette hypothèse je me baserai dans ce chapitre sur les modèles mathématiques de décision perceptive. Je commencerai donc par passer en revue les modèles de décision et de leurs bases neuronales (celles-ci ont déjà été abordées dans la section 1.6). Deux aspects souvent étudiés par ces modèles ne seront pas développés du tout ici : l'incertitude et la récompense. En effet, dans les expériences de ce mémoire, ces paramètres n'ont jamais été manipulés, nous ne nous y intéresserons donc pas. En s'inspirant de cette littérature, je proposerai ensuite un modèle simple de décision saccadique. Celui-ci me permettra d'illustrer la plausibilité de l'explication du biais vers les visages (ainsi que celui vers les animaux, voir section 3.1) par une différence de temps de traitement sensoriel.

4.2 Les modèles de décision perceptive

4.2.1 Généralités

À l'origine, les modèles mathématiques de prise de décision en psychologie ont été construits pour capturer les taux d'erreurs et les distributions de réponses dans diverses tâches de choix forcés. La plupart inclut des unités abstraites appelées accumulateurs, qui intègrent graduellement une entrée sensorielle bruitée jusqu'à ce qu'un seuil de décision soit atteint, pour ensuite déclencher une réponse comportementale. Le processus de décision, depuis la perception jusqu'à l'action, peut donc être séparé en deux phases bien distinctes :

1. Encodage du stimulus qui va permettre la représentation d'indices perceptifs.
2. Accumulation de ces indices pour atteindre un seuil de décision, puis déclencher une réponse motrice.

L'accumulation d'informations se fait, quant à elle, selon trois principes généraux :

- Les indices favorisant chaque alternative sont intégrés au cours du temps.
- Le processus est sujet à des fluctuations aléatoires.
- La décision est prise lorsqu'une alternative franchit le seuil, ou lorsque la différence entre les deux alternatives franchit le seuil.

Ce type de modèle, pourtant assez ancien en psychologie expérimentale, permet de rendre compte de l'immense majorité des résultats comportementaux (Bogacz *et al.*, 2006; Ratcliff et Rouder, 1998; Usher et McClelland, 2001; Brown et Heathcote, 2005,

2008). Ils ont de plus trouvé récemment de nombreux supports neurophysiologiques (Beck *et al.*, 2008; Bogacz *et al.*, 2009; Ferrera *et al.*, 2009; Gold et Shadlen, 2007; Hanes et Schall, 1996; Hanks *et al.*, 2006; Heekeren *et al.*, 2008; Horwitz *et al.*, 2004a; Kim et Shadlen, 1999; Lo et Wang, 2006; Ratcliff *et al.*, 2003; Roitman et Shadlen, 2002; Schall, 2001, 2002; Shadlen et Newsome, 2001, voir la section 1.6 pour plus de détails).

La phase d'encodage est rarement étudiée en tant que telle dans ces modèles. Elle est le plus souvent représentée implicitement dans les propriétés de l'accumulation (vitesse et latence), la plupart de ces études s'intéressant précisément aux mécanismes de prise de décision. Le processus d'accumulation repose quand à lui sur plusieurs paramètres qui seront présentés dans les paragraphes suivants.

Latence d'accumulation

L'encodage, le traitement purement sensoriel, prend un certain temps et retarde donc le début de l'accumulation (Ratcliff et Rouder, 1998; Usher et McClelland, 2001). Comme il a été suggéré précédemment, les accumulateurs se trouvent probablement dans la FEF, LIP ou les colliculi supérieurs. Or, ces zones nécessitent forcément quelques dizaines de millisecondes avant de commencer à recevoir de l'information, notamment de l'information pertinente pour la tâche. Dans une tâche de reconnaissance d'objets par exemple, on peut penser que ces zones se nourrissent de l'information provenant de zones sensorielles comme IT. Les neurones accumulateurs doivent donc attendre de recevoir le signal des neurones sensoriels avant de commencer à accumuler de l'information.

Niveau de départ

Une autre source de variabilité présente dans la majorité des modèles est celle sur le niveau de départ de l'accumulation (Ratcliff et Rouder, 1998; Brown et Heathcote, 2005). Comme nous le verrons par la suite, ce type de bruit est fondamental pour définir la latence des erreurs relativement aux réponses correctes. Ce bruit, sans relation avec le traitement sensoriel, pourrait être ajouté automatiquement par le système, afin de rendre la latence de la réponse imprévisible (Carpenter, 2004).

Vitesse d'accumulation

Au niveau cérébral, la qualité des indices sensoriels influence directement la vitesse d'accumulation d'information par les neurones oculomoteurs (Gold et Shadlen, 2007; Kiani *et al.*, 2008). Au niveau des modèles, ceci se traduit par la pente de l'accumulation. Plus les indices sont clairs, plus l'accumulation sera rapide. Ce paramètre va ainsi varier à chaque essai selon le stimulus (variabilité inter-essai). Il peut même varier

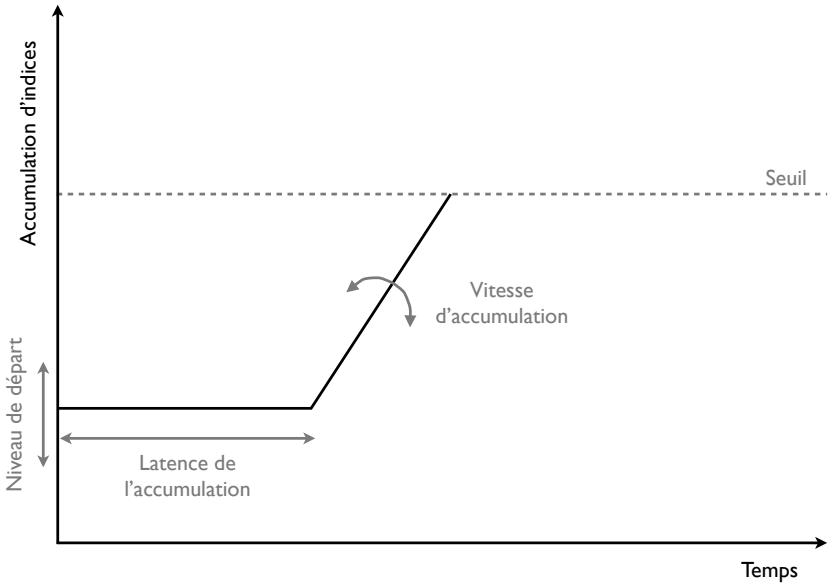

FIGURE 4.1: Schéma de l'accumulation d'information au cours du temps (cette accumulation est ici représentée sous forme de droite par mesure de simplification) et des différents bruits qui peuvent être intégrés à cette accumulation. Les trois types de bruits, ainsi que le niveau de seuil à atteindre sont représentés en grisé.

durant un même essai (variabilité intra-essai) selon les modèles. Ce facteur est une des sources de variabilité principale pour les temps de réaction. Il est important d'ajouter que dans une grande majorité des modèles de décision, cette accumulation d'information se fait de façon linéaire, ce qui n'est clairement pas le cas en réalité. En effet, la plupart des neurones visuels ayant des réponses de type phasique (très intense au début puis décroissante) et des latences variables, leur influence sur l'accumulation sera donc plus complexe qu'un simple accumulation linéaire.

Pour récapituler, trois paramètres de bruit varient donc à chaque essai :

- Latence de l'accumulation
- Niveau de départ
- Vitesse d'accumulation

Fuite

Une fuite dans l'accumulation (biologiquement probable) est souvent intégrée aux différents modèles. Celle-ci est cependant toujours fixe entre les différents essais. Elle n'a pas d'intérêt précis pour la modélisation des distributions de temps de réaction, en dehors de rendre le modèle plus plausible biologiquement.

Compétition et décision

Les paramètres que l'on vient de présenter concernent l'accumulation d'indices. Ceci se fait pour l'ensemble des choix possibles dans une tâche donnée. Il est donc nécessaire ensuite de choisir entre ces différentes alternatives. Par exemple, dans un choix à deux alternatives, on accumulera les indices en faveur de chacune, puis elles entreront ensuite dans un processus de compétition. Trois types sont généralement utilisés dans la littérature :

Les modèles de course (race model) : Les indices pour chaque réponse sont accumulés indépendamment, le premier à atteindre le seuil est le gagnant (Brown et Heathcote, 2005).

Les modèles de diffusion (drift diffusion model ou random walk) : Les indices en faveur d'une réponse compte aussi comme un indice en défaveur de la réponse alternative (Ratcliff et Rouder, 1998).

Les modèles de compétition inhibitrice : Similaires aux modèles de diffusion, mais plus les indices en faveur d'une réponse s'accumulent, plus ils inhibent la réponse alternative (Usher et McClelland, 2001).

Balance précision/vitesse

Une base de tous ces modèles est qu'ils considèrent la balance vitesse-précision (Bogacz *et al.*, 2009) comme provenant de la distance entre le niveau de départ et le seuil de déclenchement de l'accumulateur. Si la distance est faible, les sujets répondront rapidement mais avec beaucoup d'erreurs (la réponse incorrecte a plus de chance de passer le seuil avant la réponse correcte). Si la distance est grande, ils répondront lentement mais avec un très bon taux de réussite (plus de temps ici pour intégrer les informations sensorielles pertinentes). En d'autres mots, plus cet écart est grand, moins la réponse va être influencée par le bruit, et plus elle va être influencée par l'accumulation d'indices. Il est important de noter que l'important ici est la distance entre le niveau de départ et le niveau de seuil, la valeur absolue de chacun n'ayant pas d'intérêt en elle-même. Une étude électrophysiologique récente montrerait que la balance précision/vitesse serait plutôt codée via une modification du niveau de base d'activité (Bogacz *et al.*, 2009).

Nous avons donc vu que les modèles d'accumulation d'informations établis mathématiquement semblaient trouver des corrélats neuronaux directs dans des aires comme LIP, FEF ou les colliculi supérieurs. Il est important de remarquer que les mécanismes neuronaux sous-jacents à ces mécanismes pourraient très bien prendre d'autres formes que celles directement présentées ici. Ancrés dans la théorie des systèmes dynamiques, certains auteurs assimilent par exemple la décision à un basculement vers un attracteur

(Braun et Mattia, 2010). Cependant, en plus de bien reproduire les données comportementales et neurophysiologiques, les modèles d'accumulation vers un seuil peuvent aussi être considérés comme un niveau d'observation supérieur, et rendre compte des mêmes phénomènes que les systèmes dynamiques. Dans ce cadre, ils seraient alors une "sur-couche" aux comportements dynamiques d'assemblées de neurones.

4.2.2 Deux exemples concrets

Le modèle classique : diffusion de Ratcliff

Le modèle de diffusion de Ratcliff est le modèle de décision perceptive le plus connu et le plus utilisé (Ratcliff, 1978). Selon celui-ci, les décisions sont prises par une accumulation stochastique au cours du temps d'indices bruités vers un seuil de décision. La vitesse d'accumulation est déterminée par la qualité des indices perceptifs. La modification du seuil de décision affecte la balance entre précision et vitesse. Le temps de réaction correspond alors au temps requis pour atteindre le seuil de décision, ajouté au temps pour les traitements non-décisionnels (sensoriels et moteurs). Les sources de variabilité se retrouvent dans la vitesse d'accumulation (variabilité inter et intra-essai), le niveau de départ, et le temps non-décisionnel. Ces différents facteurs permettent à ce modèle de rendre compte d'un grand nombre de types de distribution de temps de réaction.

Ce modèle est adapté à des décisions à deux alternatives (2AFC) qui sont différentes du protocole de choix saccadique étudié dans cette thèse. Dans une tâche 2AFC classique, un stimulus est présenté au sujet, et celui-ci doit choisir entre deux alternatives possibles et appuyer (par exemple) sur le bouton correspondant. Dans notre cas, la tâche est sensiblement différente même si elle est encore considérée comme une 2AFC. Le sujet doit ici intégrer en même temps deux stimuli et prendre une décision basée sur ces deux. On pourrait donc imaginer ici, si l'on désirait appliquer directement le modèle de Ratcliff à nos résultats, que deux modules différents seraient nécessaires (un pour chaque image) et que c'est le premier à atteindre le seuil qui déclencherait la saccade.

Un modèle simplifié : le LBA

Un grand nombre de tentatives de simplification du modèle de Ratcliff ont été faites, comme par exemple le modèle LATER de Carpenter (Reddi *et al.*, 2003). J'ai choisi de présenter ici un exemple de simplification extrême, le modèle d'accumulateur balistique linéaire (LBA) développé par Brown et ses collaborateurs (Brown et Heathcote, 2005, 2008).

Le modèle d'accumulateur balistique supprime la variabilité intra-essai du processus d'accumulation d'évidence classique (Brown et Heathcote, 2005). Malgré cette simplifi-

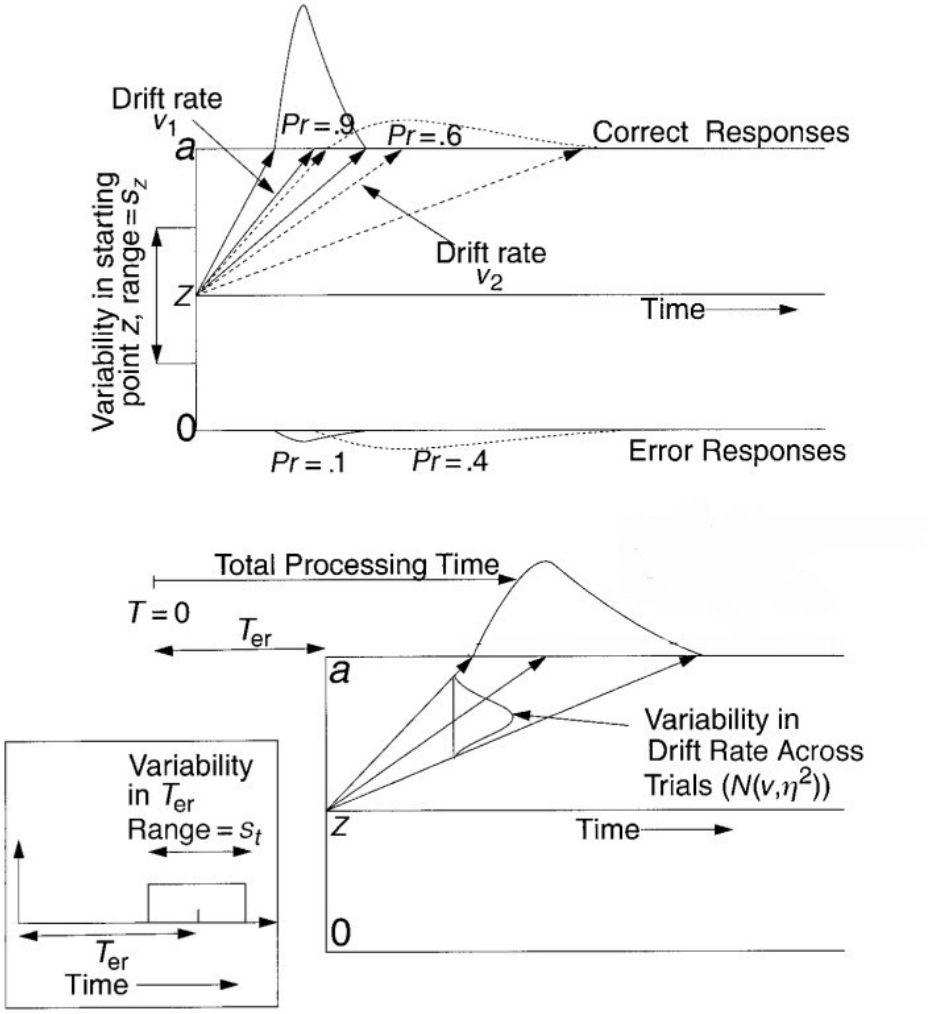

FIGURE 4.2: Une illustration du modèle de diffusion de Ratcliff et de ses paramètres. La figure du haut illustre la variabilité sur le niveau de départ, ainsi que l'influence de 2 différentes vitesses d'accumulation (v1 et v2) sur la forme des distributions de réponses correctes et incorrectes. La figure du bas illustre la variabilité de la vitesse d'accumulation entre les essais. En bas à gauche, une illustration de la variabilité sur le temps non-décisionel. Adapté de Ratcliff et Tuerlinckx (2002).

cation, il s'accommode de la plupart des phénomènes expérimentaux importants, tout en n'utilisant que quatre types de processus.

Deux processus linéaires :

- Variabilité inter-essai sur le seuil de départ
- Variabilité inter-essai sur la vitesse d'accumulation

Deux processus non-linéaires :

- Une fuite dans l'accumulation
- Inhibition mutuelle entre les alternatives

Ces mêmes auteurs ont été encore plus loin en proposant par la suite une nouvelle simplification de leur modèle : le modèle d'accumulateur balistique linéaire (Brown et

Heathcote 2008, voir figure 4.3). Celui-ci, abrégé LBA, fait abstraction de toutes les non-linéarités du modèle précédent. Bien que biologiquement plausible, la fuite et l'inhibition entre les réponses ont donc été supprimées. Tout en étant bien plus simple que la plupart des autres modèles, le LBA continue de pouvoir expliquer la plupart des phénomènes empiriques (notamment tout type de relation entre réponses correctes et incorrectes).

Pour résumer, le LBA utilise donc :

- Des accumulateurs de réponses linéaires et indépendants
- Un processus d'accumulation linéaire et déterministe

Ces modèles ont été crées pour expliquer des résultats expérimentaux composés de temps de réaction plutôt lents. Ceci explique l'ajout dans le modèle séquentiel de diffusion de processus stochastique lors de l'accumulation. Ceux-ci correspondent à une fluctuation aléatoire à chaque pas de temps de la quantité d'évidence accumulée pour chaque alternative. Ce processus permettait d'allonger les TR du modèle. Dans le LBA, ainsi que dans le modèle que nous proposerons plus tard, nous n'avons pas pour but de ralentir les réponses par l'ajout d'un nouveau processus, au contraire.

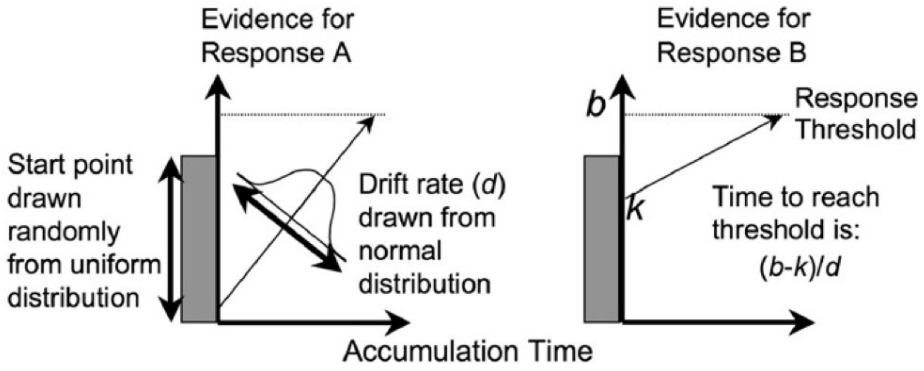

FIGURE 4.3: Une version à deux alternatives du modèle LBA. Les indices pour la réponse A sont traités dans l'accumulateur de gauche, ceux pour B dans l'accumulateur de droite. Les valeurs de départ pour le processus d'accumulation sont tirées au hasard et de façon indépendante de distributions uniformes identiques. La vitesse d'accumulation est tirée indépendamment pour chaque réponse dans une distribution normale. Une réponse est déclenchée quand le premier accumulateur atteint le seuil. Adapté de Brown et Heathcote (2008).

4.3 Vers un modèle de décision ultra-rapide

Nous venons donc de voir qu'une grande variété de modèles de prise de décision perceptive existent dans la littérature. Que peuvent-ils apporter à l'analyse des résultats présentés dans ce manuscrit ? Cette question m'a tout d'abord amené à essayer d'appliquer le modèle de diffusion de Ratcliff aux résultats des expériences présentées dans

ce mémoire. Malheureusement, ce modèle a été construit pour expliquer des protocoles similaires (2AFC) mais cependant assez différents pour être difficilement applicable ici. En effet, dans la tâche de choix saccadique, on présente deux images en même temps au sujet. Le sujet n'est donc pas en train d'intégrer une information bruitée, puis d'essayer de faire un choix parmi deux alternatives pour classer cette information. Au contraire, il est en train d'intégrer deux informations bruitées et doit choisir laquelle correspond le plus à la réponse qui lui est demandée. Le modèle de Ratcliff est donc difficilement applicable à ce design, et les différentes analyses que j'ai pu mener dans ce cadre n'ont rien amené de concluant. C'est pourquoi j'ai donc choisi de développer un modèle particulier pour la tâche de choix saccadique. Celui-ci sera bien évidemment bien moins développé au niveau analytique que le modèle de Ratcliff et n'a pas prétention à être applicable à d'autres type de données expérimentales.

Si l'on observe précisément les distributions de TR dans l'expérience 2 de l'article 1 (voir pour rappel la Figure 4.4), à la fois pour les réponses correctes et les réponses incorrectes, on voit que dans la tâche visage, les réponses correctes et incorrectes ont le même mode. On peut donc faire l'hypothèse qu'elles impliquent les mêmes processus. On partira donc du principe que les erreurs de la tâche visage sont guidées par des indices visages activés de façon erronée par l'image du véhicule. Au contraire, dans la tâche véhicule, on voit que les distributions de réponses correctes et incorrectes n'ont clairement pas le même mode, le mode des erreurs étant au contraire le même que celui dans la tâche visage. On peut donc supposer que les réponses correctes sont guidées par les indices véhicules de la cible, alors que les réponses incorrectes sont guidées par les indices visages du distracteur qui enclenche une réponse même si celle-ci ne correspond pas à la tâche (réponse automatique). Ce patron de résultat semble fondamental dans ces résultats ; l'objectif de notre modèle sera donc de retrouver ce patron de résultats bien précis.

Un autre fait marquant dans nos résultats expérimentaux est que le temps de réaction ne semble pas varier selon la difficulté de la tâche pour une même catégorie cible (voir l'étude sur le niveau de catégorisation, section 3.2), mais plutôt selon la catégorie cible (section 2.2). Ces résultats, associés au fait que les latences des populations de neurones tendent à varier selon la catégorie d'objet à laquelle elles sont sélectives (Kiani *et al.*, 2005), tendent à suggérer que le temps de réaction est principalement guidé par la latence de la population de neurone associée à la catégorie cible. Ce phénomène se traduirait par le temps de traitement sensoriel dans les modèles de décision classiques.

La latence des neurones sélectifs aux catégories de nos expériences pourrait donc être la clé pour expliquer nos résultats. Afin de supporter cette hypothèse, j'ai essayer de reproduire les distributions de temps de réaction obtenues expérimentalement dans l'expérience 2 de l'article 1, à partir d'un modèle simple de décision qui sera uniquement basé sur des différences de latences des réponses sensorielles.

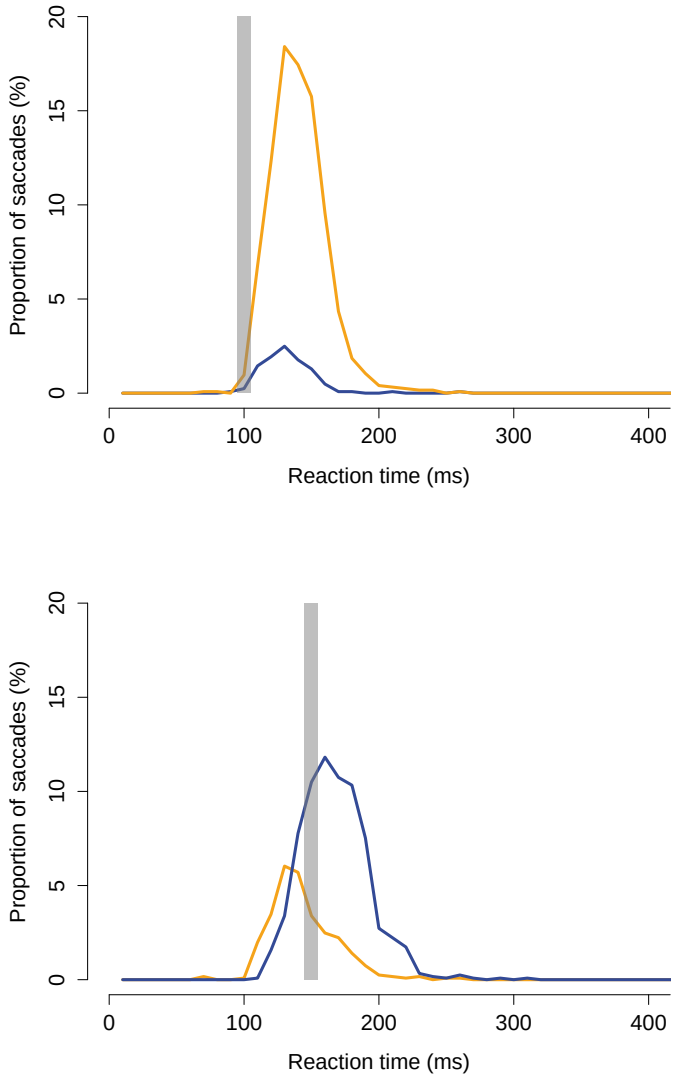

FIGURE 4.4: **Rappel des résultats de l'article 1, expérience 2.** En haut, distributions des temps de réaction lorsque la tâche est d'aller vers les visages, en bas, lorsque la tâche est d'aller vers les véhicules. Les saccades vers les visages sont représentées en orange, celles vers les véhicules en bleu.

4.3.1 Les composantes du modèle

Avant de décrire les différentes composantes inclues dans ce modèle, je vais commencer par en dresser une description générale. Puisque l'on a deux images présentées aux sujets, alors on aura deux processus d'accumulation. Chacun de ces processus d'accumulation sera nourri par deux populations de neurones sensoriels (pour les visages et pour les véhicules ici). Une compétition aura ensuite lieu entre ces deux accumulations, la première à atteindre le seuil de décision déterminera la réponse donnée, et le moment où elle a atteint ce seuil sera le temps de réaction. Différents bruits ajoutés à ces différents processus créeront alors la variabilité entre les essais. En simulant ensuite

un grand nombre d'essai, on aura ainsi des distributions de temps de réaction pour les réponses correctes et incorrectes dans chaque tâche. Ces distributions pourront être comparées aux données expérimentales.

Réponse sensorielle

Plutôt que de seulement s'intéresser au processus d'accumulation et puisque les caractéristiques de la réponse sensorielle sont au centre de cette étude, je vais me baser sur une réponse sensorielle théorique. Les neurones sensorielles peuvent schématiquement être classés en deux types selon la façon dont ils s'activent. Celles-ci sont les réponses de type phasique et les réponses de type tonique. Les neurones phasiques vont largement répondre au début de la stimulation puis rapidement devenir silencieux, alors que les neurones toniques vont avoir une réponse plus calme au début mais maintenue beaucoup plus longtemps. Pour représenter l'activité de ces deux populations combinées, j'ai choisi une fonction permettant à l'origine de décrire le mouvement dynamique d'un ressort (voir code MATLAB, section A.2). Elle permet de reproduire assez fidèlement la réponse au cours du temps d'une assemblée de neurones sensoriels composée à la fois de neurones à réponse phasique et de neurones à réponse tonique (voir la Figure 4.5 pour une illustration de sa forme caractéristique). Notre hypothèse de départ est que les neurones sélectifs à différentes catégories d'objet auraient des latences de réponse différentes. Ceci va donc constituer un paramètre important pour notre modèle. Les réponses des neurones sélectifs aux visages auront donc une latence de *FaceLatency* ms, celle des véhicules *VehicleLatency* ms.

Sélectivité des neurones sensoriels

Le système sensoriel n'est pas parfait, pour une tâche donnée (par exemple faire une saccade vers le visage), nous aurons donc quatre populations de neurones qui s'activent pour les deux images (deux par image).

Deux populations évidentes :

- Les neurones sélectifs aux traits des visages activés par l'image de visage
- Les neurones sélectifs aux traits des véhicules activés par l'image de véhicule

Mais aussi :

- Les neurones sélectifs aux traits des véhicules activés par l'image de visage
- Les neurones sélectifs aux traits des visages activés par l'image de véhicule

Bien évidemment, les neurones visages seront beaucoup plus activés par les images de visage que par les images de véhicule, ceci sera encodé dans le paramètre *preMod*, et correspond donc en quelque sorte à la "qualité" de la sélectivité des neurones senso-

riels. Plus ce paramètre est élevé, plus les réponses des neurones sont sélectives à une catégorie donnée.

Modulation par la tâche

Les paramètres doivent changer selon que le sujet doit faire des saccades vers les visages ou vers les véhicules. Comme nous l'avons vu précédemment, cette intégration des paramètres de la tâche semble se faire directement au niveau sensoriel (Ferrera *et al.*, 2009), via une modulation de la réponse de ces neurones. Par exemple, lorsque les sujets recherchent des visages, ils pourraient augmenter l'activité des neurones visages et/ou réduire celle associée aux autres objets. Ceci va donc se traduire dans notre modèle par une modification du gain de la réponse des populations de neurones sélectifs à l'objet d'intérêt. La modulation consiste ici en un simple facteur multiplicatif de la réponse, nommé *modulation*.

Cette modulation commence dans notre modèle dès le début de la réponse. Cependant, comme nous l'avons vu, nos résultats suggèrent que les saccades les plus rapides seraient plus "automatiques" que les réponses tardives et donc moins modulées par la tâche, ce qui est aussi confirmé électrophysiologiquement par des enregistrements unitaires (Treue, 2001; Maunsell et Treue, 2006). Ce délai pourrait par exemple provenir du temps nécessaire pour que le feedback se mettre en œuvre. Il serait donc intéressant de développer ce modèle avec un nouveau paramètre que serait la latence de la modulation.

Accumulation des indices

Les réponses sensorielles sont donc bien définies. Il faut maintenant intégrer ces réponses. Pour ceci, l'étape suivante consiste à simplement cumuler la réponse des neurones sensoriels dans le temps. Une fuite n'étant pas essentielle, elle ne sera pas ajoutée au modèle par mesure de simplification.

Les bruits

Afin de rendre compte de la variabilité des latences de réponse ainsi que du taux d'erreurs, du bruit doit être ajouté à ces différents mécanismes. Dans notre modèle nous avons choisi d'intégrer deux types de bruits classiques, qui constitueront donc deux sources différentes de variation pour l'atteinte du seuil :

- vitesse d'accumulation, bruit sensoriel (multiplicatif, normale, $\mu = 1$, $\sigma = Ns$)
- niveau de départ, bruit moteur (additif, uniforme, $\mu = 0$, range = [$-Nm$: Nm])

Généralement, l'aspect le plus complexe à prendre en compte pour ces modèles est la latence des distributions de réponses incorrectes. Celles-ci sont pourtant fondamentales si l'on veut modéliser de façon correcte les processus mis en jeu, la bonne prise en

compte de ce facteur constitue donc un critère important pour l'évaluation des différents modèles (Ratcliff et Rouder, 1998). Si l'on enlève les variabilités inter-essais sur la vitesse d'accumulation ou le niveau de départ, les temps de réaction pour les réponses correctes et les réponses incorrectes sont identiques. Au contraire, expérimentalement, on considère généralement que quand le choix est facile, et que les sujets doivent répondre vite, les erreurs sont plus rapides que les réponses correctes. À l'opposé, quand le choix est difficile, et que l'accent est mis sur la précision, les erreurs ont tendance à être plus lentes que les réponses correctes. Chaque source de variabilité aura des effets différents sur la forme des distributions. On considère généralement qu'une plus grande variabilité sur la vitesse d'accumulation aura tendance à ralentir les erreurs face aux réponses correctes. Au contraire, une variabilité sur le niveau de départ aura tendance à rendre les erreurs plus précoces (Ratcliff et Tuerlinckx, 2002).

Ici, la vitesse de l'accumulation est d'abord définie par la réponse sensorielle. Un bruit est rajouté ensuite sur la pente de l'accumulation (il peut être positif ou négatif, et est indépendant entre les deux alternatives). Le bruit sur le niveau de départ peut généralement être assimilé à un départ précoce de l'accumulation d'information durant le gap (Stanford *et al.*, 2010). Dans le modèle de Stanford, les deux alternatives commencent au même niveau mais l'activité des variables de décision va augmenter durant le gap (alors qu'il n'y a pourtant pas d'information à accumuler) de façon aléatoire et différente pour les deux alternatives. Ce phénomène pourrait expliquer l'accélération des temps de réaction avec l'introduction d'un gap, phénomène bien connu dans les tâches saccadiques. Dans nos expériences, l'intervalle de gap étant fixe, une manipulation du même type aurait peu d'intérêt. Puisque ce bruit est principalement moteur, j'ai décidé d'y ajouter un particularité en considérant que ce bruit symbolisait une "tendance" du sujet à favoriser à chaque essai un côté plutôt qu'un autre. Ce bruit sera donc aléatoire à chaque essai, mais pas indépendant entre les deux alternatives, si pour l'essai t, on a un avantage vers l'image de gauche, alors on aura dans le même temps un désavantage équivalent pour l'image de droite.

Seuil de déclenchement

Le niveau que doit atteindre l'accumulateur pour déclencher une réponse. Celui-ci correspond au paramètre St et sera identique entre les conditions. En effet, dans la tâche véhicule, les erreurs vers les visages ont le même mode que les réponses correctes dans la tâche visage. Le seuil de déclenchement n'a donc pas été modifié. On aurait pu penser que la plus grand difficulté de la tâche inciterait les sujets à rehausser leur seuil de façon à faire moins d'erreurs, il ne semble pas que ce soit le cas dans nos résultats.

4.3.2 Compétition entre les alternatives

Nous avons donc maintenant défini comment fonctionnaient la réponse sensorielle et son accumulation dans notre modèle. Il faut maintenant décrire comment ces différentes réponses vont être mises en compétition pour que le modèle décide quelle alternative choisir.

Une première étape consiste à sommer les réponses provoquées par chaque image. Puisque nous considérons ici que toute la modulation par la tâche se fait sur les réponses sensorielles, les réponses (visage et véhicule) pour une même image vont être sommées. Les activités associée à chaque image deviennent donc les participants à la compétition. Chaque accumulateur va ainsi concourir pour être le premier à atteindre le seuil à chaque essai, le gagnant déterminant alors la réponse comportementale. On aura donc, pour chaque essai, le temps de réaction (le moment où la première réponse atteint le seuil) et le fait que la réponse soit correcte ou non (saccade initiée du côté de la cible ou pas).

Inhibition mutuelle

Comme nous l'avons déjà évoqué, le temps de réaction tend à ne pas être modifié par la difficulté de la tâche. De plus, il a été montré que dans la tâche de choix saccadique, l'augmentation du nombre de distracteurs n'avait quasiment aucun effet sur les performances dans une tâche de détection d'animal (Drewes *et al.*, 2009). Ces résultats sont donc des arguments contre l'existence d'une inhibition mutuelle entre les deux images. Si un tel processus existait ici, une augmentation de la difficulté entraînerait forcément un ralentissement des temps de réaction. Nous faisons donc le choix de ne pas inclure ce type d'inhibition entre les deux alternatives.

Cependant, il est tout à fait possible qu'un processus d'inhibition existe entre les populations de neurones s'activant pour une même image. Comme nous l'avons vu, une image de visage va activer une majorité de neurones visages, mais aussi des neurones véhicules. On peut facilement imaginer un processus d'inhibition ici. J'ai donc intégré au modèle un mécanisme d'inhibition entre les deux populations activées par une même image (l'inhibition appliquée à la population A étant proportionnelle à l'activation de la population B). C'est le paramètre *factorInhib*

Une fois qu'une alternative a gagné la compétition, un temps fixe va s'écouler avant le déclenchement de la réponse motrice. Ceci correspond au temps nécessaire entre l'atteinte du seuil par des neurones de FEF, LIP ou SC et le déclenchement réel d'une saccade. Il est généralement considéré équivalent à 20-30 ms (Stanford *et al.*, 2010). Dans mon modèle, j'ai choisi un délai de 20 ms.

4.3.3 Présentation des résultats préliminaires

Pour tester les performances de ce modèle théorique de la tâche de choix saccadique, 2000 essais ont été simulés (le code MATLAB utilisé peut être trouvé en annexes, section A.2). Les paramètres principaux sont illustrés dans la figure 4.5.

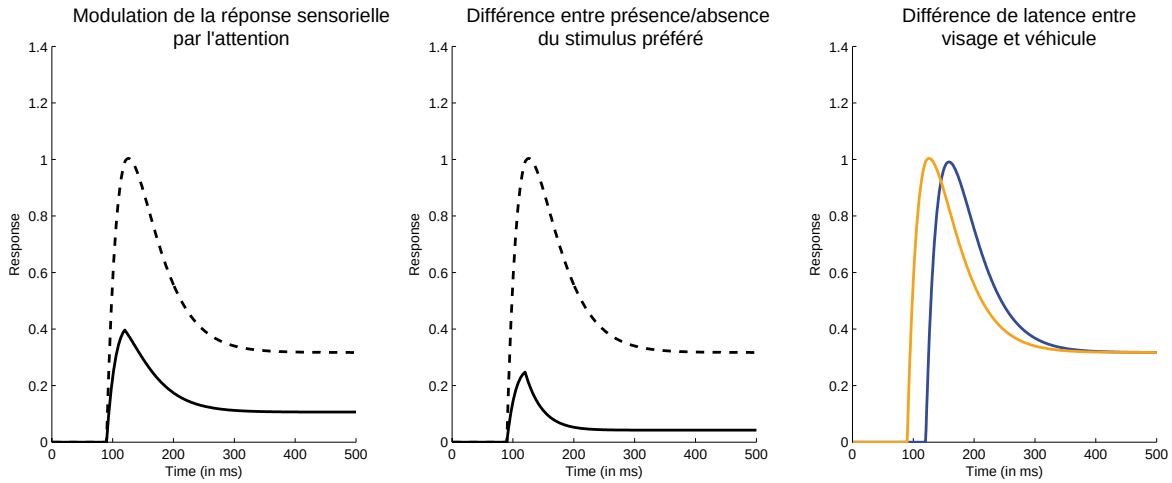

FIGURE 4.5: **Paramètres du modèle de la tâche de choix saccadique.** Modulation attentionnelle = 3. Différence entre présence absence = 4 . Différence de latence = 30 ms

L'objectif ici n'était que de prouver la plausibilité d'une explication de nos résultats par une différence de latence dans les réponses sensorielles. Comme on peut le voir dans la Figure 4.6, cet objectif est rempli. La seule différence inclue dans le modèle entre les catégories visages et véhicules est une différence de latence de la réponse sensorielle, et le patron de réponse observé dans les deux tâches est largement similaire à celui observé expérimentalement. Plus précisément, l'aspect le plus fondamental était la dynamique des réponses correctes vis-à-vis des réponses incorrectes. Dans la tâche visage, les erreurs devaient avoir les mêmes latences que les réponses correctes. Dans la tâche véhicule, les erreurs devaient avoir les mêmes latences que les réponses correctes et incorrectes de la tâche visage. C'est bien ce patron que l'on retrouve dans les résultats simulés du modèle.

Les résultats simulés ne correspondent cependant pas parfaitement aux résultats expérimentaux. Comme on peut le voir, les taux de réponses correctes ne sont pas les mêmes (expérience : 92,5% et 76,7% ; modèle : 76% et 56%), même si le plus important est que la différence entre les deux soit du même ordre. De même, les temps de réaction moyens des réponses correctes ne sont pas exactement les mêmes. Afin d'ajuster les paramètres pour correspondre aux résultats expérimentaux de façon plus précise, il faudrait utiliser une procédure mathématique d'optimisation de paramètres (comme par exemple SIMPLEX, souvent utilisé pour ajuster ce type de modèle).

En s'inspirant des modèles classiques comme celui de Ratcliff, il serait aussi extrêmement intéressant d'utiliser cette base pour réellement tester l'influence des différents

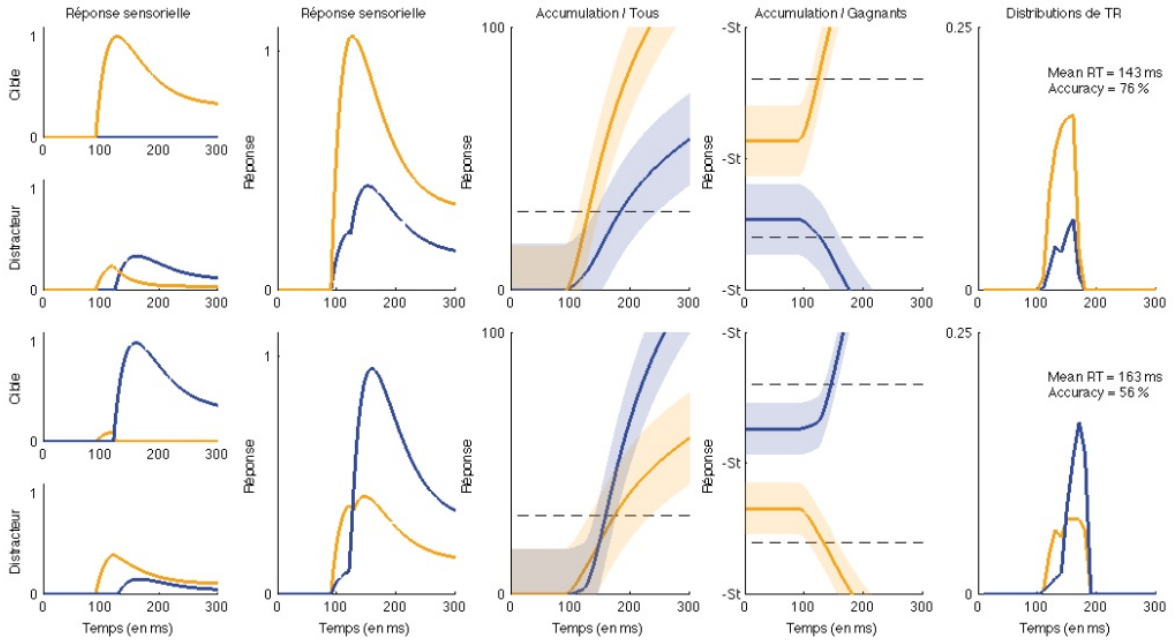

FIGURE 4.6: Résultats du modèle sur 2000 essais simulés. En haut, simulation lorsque la tâche est de faire une saccade vers le visage. En bas, lorsque la tâche est de faire une saccade vers le véhicule. **Réponse sensorielle** des différentes populations de neurones sélectifs aux deux catégories. **Réponse sensorielle** combinée entre les différents types de neurones, on a ainsi une activation pour chaque image. (cible et distracteur). **Accumulation / Tous :** courbes cumulées moyennes sur l'ensemble des essais simulés (s.e.d. en transparence). **Accumulation / Gagnants :** courbes cumulées moyennes seulement sur les essais où l'alternative a gagné. Et enfin à l'extrême droite, les **Distributions de temps de réaction** simulées par le modèle. En orange les saccades vers les visages, en bleu les saccades vers les véhicules.

paramètres pour expliquer les résultats de la tâche de choix saccadique. On pourrait par exemple tester lequel d'entre eux (parmi par exemple le temps de traitement, le niveau de seuil, la vitesse d'accumulation) permet de rendre compte au mieux des variations entre les conditions selon les manipulations réalisées dans chaque expérience.

Cependant, l'objectif principal n'était pas de réaliser un modèle qui simulerait parfaitement tous les aspects de la tâche de choix saccadique. Il était plus modestement de démontrer par une simulation avec des paramètres bien définis qu'une simple différence de latence entre deux populations de neurones pouvait être à la base des résultats bien particuliers observés dans la tâche de choix saccadique. Comme nous l'avions proposé dans l'Article 1 (section 2.2), il semble bien que ce soit le cas.

Chapitre 5

Synthèse et perspectives

Sommaire

Ces travaux de thèse ont amené plusieurs résultats importants. Je vais tenter dans ce chapitre de les intégrer dans un modèle des traitements visuels rapides, qui m'amènera à proposer une nouvelle forme de saillance prenant en compte la nature des objets et surtout l'asynchronie de leurs traitements.

5.1 Des différences de temps de traitement entre catégories

La découverte de différences de temps de traitement entre les catégories a été une surprise lors de mes premières expériences. Nous pensions à l'origine, et après l'article de Kirchner et Thorpe (Kirchner et Thorpe, 2006) que la tâche de choix saccadique allait nous permettre de chronométrer plus précisément les temps minimaux requis pour réaliser différentes tâches. Des différences entre catégories ont souvent été observées au niveau comportemental dans la littérature, dans des protocoles d'exploration oculaire (Pascalis, 1998; Pascalis *et al.*, 2002), de cécité au changement (New *et al.*, 2007), ou encore en recherche visuelle (Hershler et Hochstein, 2005). Cependant, les expériences précédentes de catégorisation rapide en scènes naturelles comparant différentes catégories nous laissaient penser qu'il n'y avait pas de raison d'observer de telles différences

entre les catégories que sont par exemple les animaux, les véhicules ou les visages (Rousselet *et al.*, 2003a; VanRullen et Thorpe, 2001). Les processus de catégorisation rapide semblaient donc invariants à la catégorie superordonnée. À notre grande surprise, nos premiers résultats ont été complètement différents de nos attentes. La sensation de "non-contrôle" que l'on ressent à chaque fois que l'on participe à ces expériences et que l'on doit faire des saccades vers les animaux et les visages trouvait une manifestation claire dans les résultats observés. En effet, les sujets sont globalement unanimes dans leurs impressions après avoir participé à une tâche de choix saccadique impliquant des animaux ou des visages : leurs yeux sont déjà posés sur la cible avant même qu'ils aient *conscience* que celle-ci était à gauche ou à droite. La grande difficulté à ne pas faire de saccades vers les visages lorsque ceux-ci étaient distracteurs faisait aussi partie des remarques fréquentes qui ressortaient des débriefings expérimentaux.

Les résultats expérimentaux ont largement confirmé les impressions des sujets : dans une tâche de choix saccadique, les différentes catégories ne sont pas traitées à la même vitesse. Ces temps de traitement variables entre catégories au niveau comportemental faisaient directement écho avec une étude en électrophysiologie parue quelques années auparavant, montrant que les neurones sélectifs à différentes catégories d'objets pourraient avoir différentes latences d'activation (Kiani *et al.*, 2005). Plus précisément, cette étude montrait, chez le singe, que les neurones codant pour les visages primates avaient des latences plus courtes que ceux sélectifs aux visages non-primates. En étendant ce résultat, on peut imaginer des différences de latences entre toutes les catégories, qui constitueraient donc une explication plausible à nos résultats (voir aussi Chapitre 4). Et si les performances dans la tâche de choix saccadique reflétaient directement les latences d'activation des neurones visuels ?

Nous avons donc un résultat comportemental clair et robuste qui correspond parfaitement à des données électrophysiologiques, la tâche de choix saccadique pourrait donc permettre de mettre en évidence des effets au niveau neuronal qui seraient invisibles en utilisant d'autres protocoles amenant des réponses plus tardives. Les temps de traitement entre catégories d'objets ne sont donc pas égaux, d'où pourrait provenir cette différence ? Trois types d'explications générales peuvent être données :

- Accélérer la traversée de la voie ventrale.
- Prendre un raccourci.
- Se contenter d'une représentation accessible plus rapidement.

5.1.1 Accélérer le traitement

Il paraît difficile d'imaginer que la propagation d'information se fasse, par exemple, plus vite pour les visages que pour les véhicules. Pourtant, on peut imaginer que les neurones sélectifs aux visages, ainsi que ceux les précédant dans la hiérarchie, sélectifs

à des traits plus locaux, puissent être optimisés pour répondre plus vite. Un mécanisme simple d'apprentissage des neurones individuels pourrait justement permettre ce phénomène. La Spike-Time Dependent Plasticity (STDP) est un mécanisme d'apprentissage très simple, dérivé de la loi de Hebb, guidant la modification des poids synaptiques des entrées d'un neurone. Selon la loi de Hebb, un neurone va renforcer le poids de ses entrées les plus stimulées. La STDP étend ce principe en y intégrant une dimension temporelle : plus le spike (potentiel d'action) afférent arrivera tôt *avant* la génération du spike du neurone intégrateur, plus la connexion sera renforcée. À l'inverse, plus le spike afférent arrivera tôt *après* la génération, moins il sera renforcé. Ceci se base donc sur une règle simple : si un spike arrive juste avant la génération, alors il est largement impliqué dans celle-ci, s'il arrive après, alors il n'a pas été utile. Cette règle permet ainsi au neurone d'apprendre implicitement lesquels de ces neurones afférents sont les plus informatifs. Ce qui va particulièrement nous intéresser ici réside dans les propriétés d'un réseau de neurones mettant en œuvre cette règle.

En effet, le groupe dans lequel je travaille a développé depuis quelques années un modèle de reconnaissance d'objets où l'information qui transite entre les neurones n'est pas contenue dans leur taux de décharge mais dans leur ordre d'arrivée (Gautrais et Thorpe, 1998; Vanrullen *et al.*, 2005). Ce modèle permet de reconnaître très efficacement (et rapidement) des modèles d'objets appris au préalable. À l'origine, il nécessitait un apprentissage supervisé pour l'entrainer à reconnaitre une classe d'objets voulue. Afin de rendre l'apprentissage de ce système non-supervisé, un mécanisme de STDP y a été ajouté comme règle d'apprentissage (Guyonneau *et al.*, 2005; Masquelier et Thorpe, 2007; Masquelier *et al.*, 2009). Il suffisait alors de présenter des flux d'images au système pour que celui-ci apprenne automatiquement les régularités dans les images, et si par exemple on lui présentait un grand nombre d'images contenant un visage, alors il développait automatiquement des neurones sélectifs à des caractéristiques de ces images qui permettaient ensuite de les classer efficacement. Ce qui nous intéresse particulièrement ici est que, avec un système de ce type, les neurones sélectifs aux caractéristiques qui se répètent le plus souvent vont tendre à décharger de plus en plus tôt (Guyonneau *et al.*, 2005). Ce résultat pourrait donc constituer une première explication plausible aux temps de traitements plus courts que l'on a observé durant ces travaux de thèse pour une catégorie comme les visages. Puisque nous sommes constamment entourés par ces stimuli depuis notre plus jeune âge et à très haute fréquence tous les jours (Sinha *et al.*, 2007), les neurones codant pour ces caractéristiques ont pu progressivement raccourcir leurs délais de décharge pour atteindre des latences plus courtes que pour des objets moins communs. Il serait intéressant dans ce cadre de tester les performances de jeunes enfants, ou même de nouveaux-nés, dans un protocole similaire, afin de voir la contribution exacte de l'expérience dans les différences entre catégories.

5.1.2 Prendre un raccourci

Une explication par la fréquence d'exposition est tout à fait plausible pour expliquer l'avantage des visages. Cependant, il est difficile d'expliquer l'avantage pour les animaux dans ce même cadre. On ne peut pas dire que, dans notre vie quotidienne moderne occidentale, nous soyons exposés aux animaux à longueur de journée. Un avantage pour cette catégorie pourrait donc résider dans d'autres mécanismes, probablement archaïques (New *et al.*, 2007). Le cerveau, malgré sa très importante plasticité, contient un certain nombre de mécanismes codés "en dur". Ce type de codage, auquel on peut se référer comme un *pré-câblage* pourrait notamment mettre en jeu des structures sous-corticales (voir section 2.2.4). Celles-ci pourraient constituer une sorte de système visuel archaïque créé pour détecter rapidement les trois sortes d'objets importants pour un animal en général : les congénères, les prédateurs, et les proies. Même si son intérêt n'est plus aussi fondamental aujourd'hui, on pourrait encore trouver quelques traces de ces mécanismes dans notre système visuel actuel. Une façon de tester ce type d'hypothèses serait l'application de ce protocole chez des patients ayant des lésions cérébrales bien définies. Par exemple, il serait intéressant de connaître les performances de patients prosopagnosiques dans la tâche de détection de visage. Si ceux-ci avaient des performances correctes, alors cela démontrerait sans équivoque que les performances que nous observons peuvent être obtenues sans passer par les aires de haut-niveau comme la FFA. D'autres types de lésion, à différents niveaux du cortex

Un autre type de raccourci pourrait mettre en jeu directement les aires visuo-motrices comme FEF et LIP. Comme nous l'avons vu dans la section 1.6, ces aires sont à la base des mouvements oculaires, contiennent des neurones sensoriels, et ont des premières latences d'activation extrêmement précoces. Elles sont donc de très bon candidats. Malgré quelques résultats allant dans ce sens (Sereno et Maunsell, 1998; Peng *et al.*, 2008; Freedman et Assad, 2006, , voir aussi section 1.3.5), le fait que ces structures puissent réellement permettre de détecter des objets comme des visages et des animaux dans le champ visuel est loin d'être avéré. Des études sont en cours actuellement dans l'équipe qui devraient nous permettre d'en savoir plus. Cependant, la grande taille des champs récepteurs dans ces aires (Blatt *et al.*, 1990; Mohler *et al.*, 1973) suggèrerait, si elles étaient effectivement utilisées, que la détection ultra-rapide d'objets ne soit pas très précise au niveau spatial. Au contraire, de nouvelles expériences [1] en cours semblent montrer qu'en plus d'être ultra-rapide, ce système de détection de visage serait ultra-précis au niveau spatial. Pour tester cette aspect, au lieu d'afficher deux images à gauche et à droite de l'écran, nous avons mis en place un nouveau protocole avec seulement une grande image présentée sur tout l'écran. La tâche du sujet était toujours la même, faire une saccade le plus vite possible vers le visage, mais cette fois celui-ci

1. Mathey, M. A., Crouzet, S. M. & Thorpe, S. J. (2010). Ultra-rapid saccades to faces : the effect of target size. VSS, Naples, Florida.

pouvait apparaître dans 16 positions différentes ! Nos premiers résultats montrent que le coût temporel associé à cette multiplication des choix possibles était d'environ de 10 ms seulement. En plus de ne pas être affectées par le nombre de positions possibles, les saccades des sujets étaient étonnement précises, avec 97% des saccades orientées dans le bon cadran ($\pm 22{,}5^{\circ}$autour de la bonne direction), les saccades les plus rapides n'étant pas moins précises que les autres. Ces nouvelles données confortent l'hypothèse d'un système totalement parallèle et feedforward qui permettrait donc de détecter la présence et la position d'un visage dans le champ visuel en seulement quelques dizaines de millisecondes, et ceci de façon extrêmement robuste et précise.

Il est important de noter ici que cette conception du système visuel comme un système basé sur une multitude de voies qui fonctionnent et interagissent en parallèle rejoint directement les théories de Jean Bullier (Bullier, 2001, 2004; Bullier et Nowak, 1995). Selon lui, par opposition à la vision purement ventrale et hiérarchique, on trouverait des multitudes de voies qui auraient des dynamiques différentes. Par exemple, à l'intérieur même de la voie ventrale, on pourrait avoir la voie magnocellulaire qui permet une avancée plus rapide de l'information et pourrait donc moduler l'activité de la voie parvocellulaire (par feedback) alors même que cette voie n'a pas terminé sa première vague. Cette conception permet surtout de rendre compte des latences d'activations des différentes aires du système visuel, qui sont loin de respecter parfaitement la vision hiérarchique.

5.1.3 Utiliser une représentation précoce

Dans le règne animal, il est très courant d'utiliser des *proxy* pour résoudre des tâches complexes. Ceux-ci sont largement utilisés, notamment pour échapper aux prédateurs. On peut trouver de nombreux exemples, qui commencent par des systèmes visuels extrêmement simples uniquement dédiés à cette tâche et qui se basent donc sur une information très *brute*. Par exemple, certains mollusques ou crustacées possèdent seulement un détecteur de luminosité sur le dessus. Si la luminosité baisse soudainement, ce qui peut correspondre à un prédateur leur passant au-dessus, alors ils prennent automatiquement la fuite. Dans notre cas, la détection de visages étant primordiale au quotidien, le système a certainement trouvé des "astuces" pour le faire rapidement et automatiquement. Comme nous l'avons montré dans l'article 2, cette astuce pourrait être en partie basée sur une information bas-niveau qu'est le spectre d'amplitude dans le domaine de Fourier. Cette information, extraite de façon locale (donc se rapprochant plus de ce que peut faire un traitement en ondelettes par exemple), serait accessible très vite dans la voie visuelle et pourrait alors guider les saccades.

Cette information est plus proche d'une solution "quick & dirty" que d'une solution

fiable et robuste. C'est pourquoi elle ne serait pas utilisée lorsque l'on teste les sujets avec des réponses manuelles plus lentes. Notre principal système de reconnaissance d'objets, nous permettant une vision détaillée et précise n'a pas intérêt à se baser complètement sur ce type d'informations. Son but n'est pas forcément de permettre une réaction la plus rapide possible mais au contraire de procéder à une analyse détaillée. Il est important de noter que si ces informations "quick & dirty" sont utilisées pour l'initiation rapide de saccades, c'est justement car les saccades sont une réponse motrice très particulière qui n'entraîne aucun risque. Une erreur en saccade "coûte" à peine 200 ms, le temps d'en initier une nouvelle. À l'opposé, les réponses manuelles se doivent d'être contrôlées un minimum car elle peuvent impliquer un certain risque. Cette idée de contrôle peut aussi amener à une autre observation intéressante concernant les résultats dans la tâche de choix saccadique. En effet, comme nous l'avons vu, les saccades les plus rapides semblent relativement indépendantes de la tâche que le sujet doit effectuer, ce qui se manifeste par le nombre important de saccades précoces vers les visages alors que la tâche est de faire des saccades vers les véhicules (phénomène que l'on retrouve aussi avec le couple animal/véhicule). Cette observation fait directement écho aux enregistrements de neurones dans les études sur l'attention (notamment ici les études sur l'attention liées à la tâche). Un fait bien établi maintenant est que l'attention (qu'elle soit spatiale ou liée à la tâche) influence plutôt la composante tardive de la réponse du neurone (après 100 ms, voir Treue, 2001, pour un exemple). Dans ce cadre, la tâche de choix saccadique pourrait être un outil de choix pour observer les informations disponibles très précocement, avant même que l'attention puisse moduler efficacement la réponse.

5.2 Attention, saillance et latence

La grande difficulté à effectuer des saccades vers les véhicules lorsque les distracteurs sont des visages (ou des animaux) correspond à une attraction automatique du regard en fonction de la catégorie de l'objet. Ces observations rejoignent les questions sur le traitement pré-attentif de la catégorie de l'objet. Deux protocoles largement utilisés en psychologie expérimentale, la recherche visuelle et la double-tâche, ont mené à des résultats apparemment contradictoires concernant cette question. En effet, la détection d'un animal dans une scène pourrait se faire sans mise en place de processus attentionnels (double tâche, Li *et al.*, 2002), cependant ces mêmes animaux, dans un affichage de recherche visuelle, ne causerait pas d'effet pop-out (VanRullen *et al.*, 2004; VanRullen, 2009). VanRullen explique cette apparente contradiction en proposant deux types d'assemblages différents des traits locaux (VanRullen, 2009) : (i) un assemblage automatique et pré-câblé, qui permet de reconnaitre les objets "naturels" (que l'on retrouve dans la plupart des modèles de reconnaissance d'objets, Serre *et al.* 2007c), (ii) un as-

semblage attentionnel, qui permet de réaliser un grand nombre de tâches additionnelles (celles utilisées généralement en psychophysique, comme reconnaître le disque rouge et vert parmi des disques verts et rouges).

L'assemblage pré-câblé permettrait aux objets tels que les animaux d'être traités sans attention. Lorsqu'un traitement peut être fait de façon purement ascendante (ou feedforward), alors l'attention ne serait pas requise. Dans le cas des animaux, la première vague ascendante allant de la rétine jusqu'à IT permettrait à elle seule cette reconnaissance, selon un mode pré-câblé. Pourquoi alors les animaux ne *pop-out*-ils pas d'un affichage de recherche visuelle ? L'explication pourrait provenir des caractéristiques des neurones de IT (par exemple), très sélectifs mais ayant des champs récepteurs très larges. Dans un affichage où d'autres objets peuvent entrer en compétition à l'intérieur même du champ récepteur concerné (comme en recherche visuelle), la reconnaissance serait perturbée (concept de compétition biaisée, Desimone et Duncan, 1995). On pourrait donc reconnaître très rapidement, et en une seule vague d'activation, la catégorie de l'objet présentée (en tout cas pour des catégories génériques comme animal, visage, véhicule), mais cette capacité trouverait sa limite dès que l'affichage deviendrait trop "serré", avec d'autres distracteurs entrant en compétition au sein des mêmes champs récepteurs que ceux utilisés pour l'objet recherché.

La taille des champs récepteurs permet aussi de rendre compte de la différence entre les catégories animal et visage. Comme nous l'avons vu, il semble que le cas des visages soit sensiblement différent car ils provoqueraient un effet *pop-out* dans un affichage de recherche visuelle (Hershler et Hochstein, 2005). Cette différence pourrait aussi s'expliquer par la taille des champs récepteurs. Nous avons montré dans l'article 2 que la détection de visages pourrait être basée sur une information plus bas-niveau (voir aussi : VanRullen, 2006), probablement basée sur l'activité de neurones en amont dans la voie visuelle, comme ceux de V4. Les neurones de V4 ayant des champ récepteurs moins grands que ceux de IT, les neurones de cette aire seront donc moins sensibles à la compétition que ceux de IT. Ils seront moins sensibles dans le sens où il faudra que le distracteur soit plus rapproché pour diminuer l'activité du neurone. Dans ce cadre, les catégories qui peuvent être détectées à partir de neurones plus en amont dans la voie visuelle seront moins sensibles à la compétition à l'intérieur des champs récepteurs. Un effet *pop-out* pourrait donc être observé pour ces catégories, tant que l'écart entre les items de l'affichage reste supérieur à la taille des champs récepteurs.

Nos résultats s'accordent donc globalement très bien avec l'idée de reconnaissance ascendante pré-câblée. Quelles informations additionnelles sont donc amenées par nos résultats saccadiques ? On peut considérer que les trois catégories principalement étudiées dans les expériences de cette thèse (animal, visage, véhicule) peuvent certainement toutes les trois être traitées sans attention (cela a en tout cas été montré pour les ani-

maux et les visages). Elles seraient donc toutes trois reconnaissables en une seule vague feedforward. Le *pop-out* serait quant à lui réservé aux visages, car des caractéristiques plus bas-niveau pourraient être mises en jeu, impliquant des champs récepteurs moins large, et donc moins sujets à la compétition.

Les résultats de nos expériences, montrant des différences de temps de traitement entre les catégories elles-mêmes, permettent de préciser encore les différents phénomènes en jeu lors de la reconnaissance rapide d'objet dans les scènes naturelles. Ainsi, les différentes catégories s'organiseraient selon un *continuum* de temps de réaction, avec des visages détectés en un temps records (100-110 ms), des animaux qui prendrait un peu plus de temps (120-130 ms) et des véhicules en queue de peloton (plutôt 140-160 ms, ce qui correspond aussi aux temps enregistrés pour l'extraction des informations globales de contexte). Le pré-câblage ne serait donc pas aussi "efficace" pour toutes les catégories.

Ces temps de traitement différents semblent directement associés à une capacité de capture attentionnelle. Ainsi, les sujets avaient tendance à produire un grand nombre de saccades vers les visages, même lorsque ceux-ci étaient les distracteurs et que la tâche était de faire des saccades vers les véhicules. Ce phénomène se rapproche du phénomène généralement appelé capture attentionnelle, qui peut généralement être observé pour deux raisons : (i) la saillance très forte d'un stimulus (comme un rond rouge parmi des ronds verts Treisman et Gelade, 1980), (ii) ou de façon encore plus puissante et surtout plus difficile à contrôler par l'affichage brutal d'un nouvel objet (Yantis et Jonides, 1984). Un effet de capture attentionnelle spécifique aux visages avait déjà été mis en évidence (Bindemann *et al.*, 2007, 2005). Nos résultats confirment clairement ce phénomène, avec des saccades précoces très difficiles à inhiber dirigées vers les visages (mais aussi vers les animaux face à des véhicules). Il semble donc que le temps de traitement associé à chaque catégorie se retrouve dans la capacité de cette même catégorie à capturer l'attention. La saillance intrinsèque des catégories pourrait donc être reliée directement à leurs temps de traitement. Cette observation peut donc permettre de proposer une nouvelle composante de la saillance ascendante basée sur la catégorie de l'objet.

Vers une nouvelle composante de la saillance

Cet ensemble d'observations permet de faire émerger une nouvelle composante qui pourrait entrer en jeu dans le concept de saillance. Cette hypothèse reprend la conception classique (Itti et Koch, 2001) de saillance guidée par les indices bas-niveaux, et y ajoute une nouvelle composante basée sur le traitement asynchrone des différentes catégories d'objets.

Pour résumer, comme nous l'avons vu dans la section 1.6, la saillance des différents objets dans notre champ visuel est codée dans des cartes topographiques. Ces cartes

représentent généralement à la fois une ”carte” attentionnelle et/ou la destination des mouvements oculaires. L'ensemble des stimuli du champ visuel seraient représentés dans un premier temps dans cette carte, avant qu'elle ne se trouve modulée par la tâche à effectuer, ou l'objet que l'on recherche. Si l'on intègre à ce cadre classique l'asynchronie entre les différentes catégories, on voit apparaître un modèle cohérent des résultats que l'on a observé tout au long de cette thèse. Les objets qui auraient les latences les plus courtes seraient ainsi représentés en avance dans cette carte, ce qui leur donnerait un avantage conséquent dans la compétition. On parlerait alors ici d'une forme de saillance basée sur la catégorie d'objet. Il n'est pas question ici de remettre en cause la notion de saillance classique (Itti et Koch, 2001) basée sur les aspects bas-niveaux (couleur, orientation, luminance, mouvement), mais d'y ajouter une nouvelle composante. Si les objets naturels peuvent être traités de façon purement ascendante et très rapide, alors il n'y a pas de raison pour que la nature des objets n'entre pas en jeu dans les compétitions pour la saillance. Cette hypothèse permettrait d'expliquer à la fois la tendance des saccades durant une exploration à se rendre de façon préférentielle vers les objets (Einhäuser *et al.*, 2008b), ainsi que les meilleures performances de prédiction des modèles de saillance en y ajoutant une composante de ”détection de visage” (Cerf *et al.*, 2009; Birmingham *et al.*, 2009).

Ce principe de carte de saillance influencée par les objets pourrait s'appliquer à toutes les informations qui peuvent être extraites d'une simple vague feedforward, permettant d'arriver dans les cartes de saillance à temps pour jouer un rôle dans le choix saccadique. On peut par exemple imaginer que les mots (ou au moins certains d'entre eux, comme notre prénom), qui sont des stimuli auxquels nous sommes particulièrement exposés, puissent entrer dans cette catégorie. Il est cependant possible que seules quelques catégories d'objets (les visages certainement, les animaux très probablement) aient à leur disposition des mécanismes de traitement assez rapides pour avoir le temps de jouer un rôle ici, en tout cas pour la toute première saccade. Ceux-ci seraient ainsi ”programmés” pour aller directement influencer la carte de saillance. Ce programme serait codé en dur (pré-câblage) dans le système visuel et permettrait d'attirer l'attention et/ou le regard automatiquement vers ceux-ci. Ces mécanismes, bien qu'étant basés sur un principe extrêmement simple, pourrait jouer un rôle très important au niveau comportemental, au moins dans la sélection des objets d'intérêt dans le champ visuel. Le protocole de choix saccadique, en allant chercher l'information très précocement, serait ainsi un protocole de choix pour ”jeter un regard” aux représentations précoces de ce type.

5.3 Perspectives

Comme nous l'avons vu tout au long de cette thèse, de simples protocoles psychophysiques peuvent s'avérer extrêmement fructueux pour la compréhension des mécanismes de traitement rapide de l'information. Cependant, pour tester les hypothèses soulevées au long de cette thèse, et continuer à explorer les mécanismes de perception visuelle, l'utilisation de techniques d'imagerie cérébrale devient sans aucun doute indispensable. La rapidité des phénomènes observés rend cependant cette question très difficile à aborder. En plus de leur dynamique rapide, ces mécanismes pourraient mettre en jeu des populations de neurones très locales au niveau cérébrale. Les techniques d'imagerie actuelles ne permettent pas encore d'aborder ces deux aspects à la fois. En effet, les techniques puissantes au niveau temporel (EEG, MEG, TMS) s'avèrent limitées à l'échelle spatiale. À l'inverse, les techniques permettant une vrai investigation spatiale ne permettent pas une bonne résolution temporelle (IRMf). Ainsi, malgré le nombre immense d'études qui sont maintenant faites en utilisant l'IRM fonctionnelle en neurosciences, cette technique ne permet pas pour l'instant d'avoir réellement accès à la dynamique des traitements en dessous de la seconde. Cette échelle est pourtant fondamentale pour une vrai compréhension des mécanismes cérébraux, et encore plus pour ceux de reconnaissance rapide. Des tentatives intéressantes existent pour augmenter la résolution temporelle de cette technique, en enregistrant l'activité d'une seule coupe par exemple plutôt que de tout le cerveau, ce qui permet de réduire la durée d'un cycle d'enregistrement (Sabatinelli *et al.*, 2009). Mais cette astuce requiert un postulat fort sur les zones d'intérêt. Malgré cette innovation intéressante, l'IRMf reste donc limitée au niveau temporel. Son utilisation pour la compréhension des mécanismes mis en jeu durant ces travaux de thèse reste donc difficile.

La solution se trouve donc certainement dans la combinaison de ces différentes techniques (par exemple l'enregistrement simultané de données EEG et IRMf : Debener *et al.*, 2006). En étant enregistrée simultanément à l'EEG, l'IRM permet d'effectuer une reconstruction de source précise à partir de l'activité EEG, ce qui donne accès à un signal temporellement et spatialement intéressant. Cet ensemble de techniques, associées aux outils mathématiques de décodage multi-voxels (appelées généralement MVPA, Pereira *et al.*, 2009; Norman *et al.*, 2006; Mur *et al.*, 2009) pourront certainement permettre à terme d'avoir un réel accès aux activations cérébrales dans la voie ventrale, et surtout à leurs dynamiques. Le but de ces combinaisons de techniques étant de pouvoir à court ou moyen terme vraiment "jeter un œil" au trajet de certain type d'information dans la voie ventrale.

Cette thématique sera largement abordée durant mon stage post-doctoral qui sera effectué au sein du "Department of Cognitive and Linguistic Sciences" de Brown University sous la direction du Dr. Thomas Serre. Une partie significative du projet consistera

à essayer de faire un pas de plus dans la stimulation naturelle. Pour l'instant, l'utilisation de scènes naturelles statiques flashées soudainement sur l'écran correspond encore largement à des conditions de laboratoire très particulières. Il serait donc intéressant d'aller encore plus loin et de changer de paradigme pour essayer de tester la vision dans un cadre encore plus naturel, comme par exemple la vision d'un film en continu. En plus de ce projet, un phénomène qui m'a particulièrement intéressé durant mes travaux et que je souhaiterai développer largement dans le futur est le mécanisme de *feature-based attention* (attention basée sur les caractéristiques). En effet, cet effet de l'attention est beaucoup moins bien compris que son aspect spatial, et les mécanismes précis mis en place pour moduler l'activité du système visuel pour le rendre particulièrement "sensible" à différentes caractéristiques restent assez mal connus pour le moment.

Annexe A

Annexes

Annexe A.1: Analyse des temps de réaction

Lors de mes travaux expérimentaux durant cette thèse, j'ai toujours utilisé l'analyse des temps de réaction pour tester des hypothèses et inférer les mécanismes sous-jacents. En plus des analyses classiques sur les temps de réaction moyens (ou médians) et les pourcentages de réussite, j'ai aussi souvent utilisé ce que nous appelons les temps de réaction minimums (TRmin). Cette mesure est largement acceptée par la communauté scientifique (en témoigne les nombreuses publications y faisant référence), mais pourtant très peu utilisée en dehors de l'équipe dans laquelle je travaille. Cette partie annexe vise à justifier son utilisation et à expliquer ce qu'elle apporte en comparaison à d'autres mesures.

L'idée de départ du TRmin est de capter le moment précis où une information devient disponible au niveau comportemental. Cette mesure s'inspire des analyses de signaux type potentiels évoqués (ERP), où les essais de chaque condition sont moyennés et où l'on teste ensuite à quel moment précis le signal devient significativement différent entre la condition A et la condition B. L'analyse la plus courante est de mesurer la latence du pic de la différence (ce qui a donné lieu aux latences ERP très connues que sont la P100 ou la N170). Une autre information fondamentale est le moment précis où les deux courbes commencent à diverger. Cette dernière méthode est celle qui a été utilisée pour la différentielle à 150 ms entre cible et distracteurs dont nous avons parlé dans l'introduction de ce mémoire. Le TRmin, pour les temps de réaction, est donc une adaptation de cette dernière méthode d'analyse appliquée aux TR comportementaux.

Le problème ici étant que, pour un essai, on n'a pas de signal quantitatif au cours du temps, mais une seule et unique valeur de TR, associée au fait qu'elle soit correcte ou non. Plutôt que de moyenner les signaux entre tous les essais (comme avec l'ERP), on va donc ici grouper tous les TR pour avoir une distribution. Sur cette distribution, on calcule généralement le pourcentage de réponses correctes, ainsi que la

moyenne (ou médiane) de ces temps de réponses. Si l'on continue l'analogie avec les réponses électriques, on peut alors essayer de voir à partir de quel moment les effectifs de réponses correctes et de réponses incorrectes commencent à être significativement différentes. Pour ceci on regroupe les TR pour une condition donnée par groupes d'intervalle de temps. Par exemple, dans l'intervalle 110 ms, on comptabilisera tous les TR compris entre 105 et 114 ms, et ainsi de suite. Cette procédure, appliquée de façon différentielle entre les réponses correctes et incorrectes, permet de tracer des courbes comme celle montrée en exemple dans la figure A.1. On applique un test du χ^2 (test statistique adapté aux effectifs) sur chaque intervalle de temps indépendamment. Puis on sélectionne le premier intervalle à devenir significativement différent entre réponses correctes et incorrectes (une série de 5 intervalles différentes de suite est requise dans nos expériences) qui deviendra le TRmin.

Ce type d'analyse repose donc sur un postulat fort qui est que les différentes réponses d'une distribution de temps de réaction ne sont pas toutes régies par le même processus, avec une variabilité uniquement motrice, mais que l'information guidant les réponses est accumulée au cours du temps. Dans ce cadre, les réponses très précoces seront basées sur très peu d'information (voir aucune et le sujet sera alors au niveau chance). Plus on avancera dans le temps, plus les réponses seront correctes. C'est généralement ce que l'on observe dans nos tâches de catégorisation rapide ou ultra-rapide, avec des erreurs qui ont tendance à être plutôt précoces.

Une critique qui peut être faite est que la valeur finale de TRmin peut être guidée par un seul sujet. Cette observation est vraie, puisque l'on mélange les réponses de l'ensemble des sujets pour ne considérer l'ensemble des données comme ne provenant que d'un méta-sujet. Si par exemple un seul d'entre eux a répondu dans l'intervalle de temps X, alors la différence statistique pour cet intervalle sera entièrement guidée par ce sujet. Il est cependant important de remarquer que si ces réponses rapides sont au niveau chance (autant de correctes que d'incorrectes), alors elle n'auront pas d'effet sur le TRmin. En revanche, si ces réponses sont plutôt correctes, alors elle pourront effectivement à elles seules faire ressortir un TRmin. Mais ce dernier cas correspond précisément à ce que l'on recherche : le temps minimum pour réaliser la tâche. L'idéal serait bien sûr de calculer un TRmin individuel pour chaque sujet, malheureusement, ceci nécessiterait une quantité de données par sujet et par condition extrêmement grande.

Une autre alternative pour calculer un TRmin individuel est d'utiliser les distributions de réponses cumulées (voir figure A.1). Cependant, un nouveau problème se pose. En effet, si le sujet a produit un nombre important de réponses précoces avant d'avoir assez d'informations pour être sélectives, alors ces réponses peuvent décaler l'intervalle de temps où le test devient significatif. Ce décalage ne correspond alors plus au "vrai" moment où les réponses commencent à être sélectives.

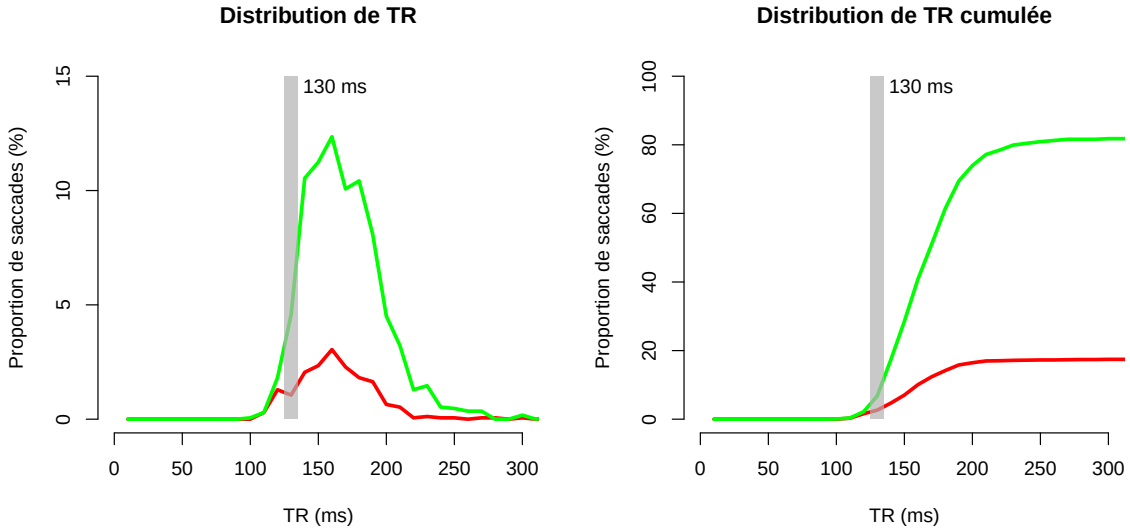

FIGURE A.1: Distribution de temps de réaction (à gauche) et la même distribution cumulée (à droite). Le cumul ne change pas le TRmin, cependant si les sujets ont produit un nombre important de réponses précoces avant d'avoir assez d'informations pour être sélectives, alors ces réponses peuvent décaler l'intervalle de temps où le test devient significatif.

Comparaison entre deux conditions

Nous avons donc vu l'analyse faite sur une condition donnée, pour déterminer le moment précis où les réponses commencent à être guidées par l'information visuelle relative à la tâche. Souvent, comme cela a été le cas dans les expériences de ce mémoire, nous cherchons à savoir à quel moment précis un effet peut se mettre en place entre deux conditions différentes. Par exemple, dans le cas de l'expérience sur l'influence du contexte sur la catégorisation d'objet, nous voulions savoir à partir de quel moment les saccades vers les animaux commençaient à être influencées par la congruence du contexte. Deux hypothèses sont ici possibles, soit les processus qui se mettent en place le font dans un temps absolu (arrivé de l'effet 150 ms après la présentation des images pour tous les sujets) ou en temps relatif (arrivé de l'effet 70 ms après les premières réponses de chaque sujet). Dans le premier cas, le TRmin est un bon moyen de s'affranchir des différences inter-sujets qui deviennent alors du bruit. Dans le deuxième cas, il est plus adapté d'avoir recours à un procédé appelé la vincentisation (Ratcliff, 1979). Celle-ci permet de regrouper les réponses des différents sujets en s'affranchissant des différences globales de TR ente eux. Par exemple, le premier quantile de réponses du sujet 1 sera combiné avec le premier quantile de réponses du sujet 2, même si les TR entre ces deux sujets sont différents. Cette méthode est particulièrement adaptée pour combiner les distributions de différents sujets car la distribution finale aura les mêmes caractéristiques que les distributions individuelles (Ratcliff, 1979). Cependant, si un effet apparaît, comme nous l'avons vu précédemment, selon un temps absolu, cette

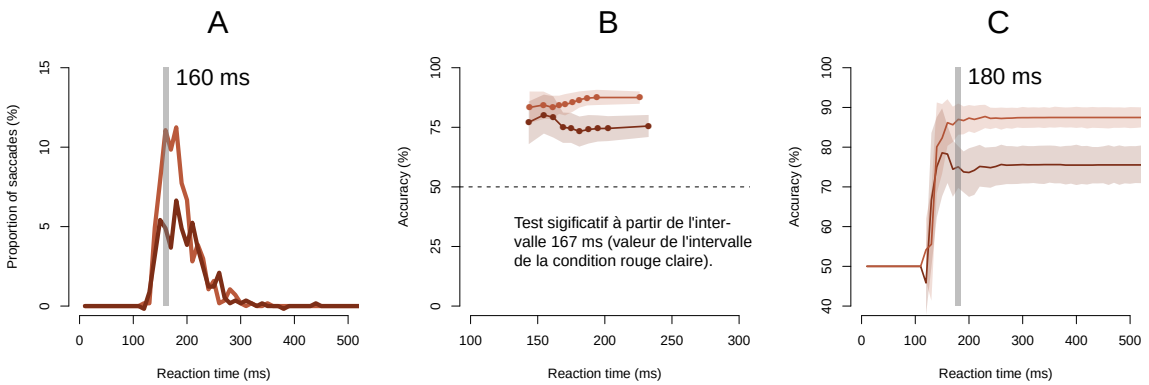

FIGURE A.2: Différentes méthodes pour mesurer le moment précis où 2 conditions se séparent. (A) Précision cumulée moyennée inter-sujets. (B) Précision vincentisée et cumulée, toujours inter-sujets. (C) "TRmin". Comme on peut le voir, les deux premières méthodes décalent l'apparition de l'effet de 20 ms du fait qu'elles sont basées sur des réponses cumulées. Avec un nombre plus important de données par sujet permettant de se passer de cumulation, ces méthodes s'avèreraient très intéressantes. La différence principale entre A et B étant que dans A on suppose que l'effet apparaît de façon absolue, alors que dans B, il apparaît de manière relative à la distribution de réponses de chaque sujet.

procédure aura tendance à l'écraser.

Dans l'expérience précédemment citée, nous avions essayé différentes méthodes pour tester le moment d'apparition précis de l'effet congruence. Celles-ci sont illustrées dans la figure A.2. Les 3 méthodes présentées sont :

- Un "TRmin" entre les distributions de la condition congruente (réponses correctes moins réponses incorrectes) et la distribution de la condition incongruente (Figure A.2, A).
- Après vincentisation des réponses, nous calculions la précision dans chaque quantile pour chaque sujet et chaque condition. Nous faisions ensuite la moyenne de ces valeurs (précision moyenne et TR moyen) pour chaque quantile. (Figure A.2, B).
- Un calcul de la précision sur les intervalles de temps cumulés pour chaque sujet et chaque condition, puis un test t (corrigé pour comparaison multiple) entre les deux courbes obtenues (Figure A.2, C).

L'analyse des temps de réaction par une approche temporelle est donc constamment une balance entre précision et nombre de données à disposition pour chaque sujet. Les approches présentées ici seraient pour la plupart plus fiables au niveau statistique si elles pouvaient être utilisées sur des données avec un très grand nombre d'essais par sujet et par condition. Dans tous les autres cas, la mesure de TRmin reste une mesure extrêmement intéressante pour l'analyse temporelle de l'accumulation d'information.

Pourcentage correct ou d' ?

Pour mesurer la précision des sujets dans les différentes expériences de ce mémoire, je me suis toujours basé sur le pourcentage de réponses correctes. En psychophysique, cette mesure est souvent remplacée par une mesure de discriminabilité issue de la théorie de la détection du signal, le *d'*. Ceci a été un choix délibéré car dans le cas précis de la tâche de choix saccadique, deux images sont présentées au sujet, l'une d'elles étant toujours une cible. Ses réponses ne peuvent donc être que de deux types : HIT (réponse cible sur une cible) et des FA (False Alarm, réponse distracteur sur une cible). Ici, contrairement à d'autres tâches, comme par exemple le go/no-go, les sujets ne peuvent pas produire de MISS (réponse distracteur sur une cible) ni de CR (Correct Rejection, réponse distracteur sur un distracteur). La somme des taux de HIT et de FA est donc toujours égale à 1, et le d' est donc toujours parfaitement corrélé au pourcentage correct. Ce dernier étant plus aisément compréhensible et apportant autant d'information dans notre cas, j'ai donc fait le choix de toujours utiliser celui-ci.

Annexe A.2: Code MATLAB pour le modèle de décision

A.2.1 Script principal : Modèle

```
param;
nbtrial=2000;

% PARAM RESPONSE %(should not play with)
t=1:1:500; %time
A=0.1; % env amplitude
B=-0.5; % abs(B) = sustained level
w0=0.03; % env largeur de la courbe

% Response gain
FaceGain = 1;
VehicleGain = 1;

% FEATURE BASED ATTENTION
% gain modulation / global gain increase by attention
mini = 1; attentionLatency = 2; % in ms
maxi= 2.5; tmax = 3; % amplitude max and latency of the max in ms

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% INIT %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
RFTarget_trial = zeros(nbtrial,length(t));
RFDistra_trial = zeros(nbtrial,length(t));
RVTarget_trial = zeros(nbtrial,length(t));
RVDistra_trial = zeros(nbtrial,length(t));
```

```
RTF = zeros(nbtrial,1);
correctF = zeros(nbtrial,1);
RFTarget_trialNm = zeros(nbtrial,length(t));
RFDistra_trialNm = zeros(nbtrial,length(t));
S0_F = zeros(nbtrial,1);

RFCTarget_trial = zeros(nbtrial,length(t));
RFCDistra_trial = zeros(nbtrial,length(t));
RVCTarget_trial = zeros(nbtrial,length(t));
RVCDistra_trial = zeros(nbtrial,length(t));

RTV = zeros(nbtrial,1);
correctV = zeros(nbtrial,1);
RVTarget_trialNm = zeros(nbtrial,length(t));
RVDistra_trialNm = zeros(nbtrial,length(t));
S0_V = zeros(nbtrial,1);

% LEXIQUE :
% RF = Response Face
% RV = Response Vehicle
% a = absent
% p = present
% A = attended

for i = 1:nbtrial

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %%%%%%%%%%%%%%%%%%%%%%%%% CREATE THE MODEL %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    %%% Calculating the latency noise for each pop
    latN_RFp = round(random('unif', -Nl, Nl));
    latN_RVa = round(random('unif', -Nl, Nl));
    latN_RVp = round(random('unif', -Nl, Nl));
    latN_RFa = round(random('unif', -Nl, Nl));

    %%% Creation of the neural responses for each population
    %present
    RFp_trial(i,:) = ...
        NeuralPopResponse(t,A,B,w0,FaceLatency + latN_RFp,FaceGain);
    %absent
    RVa_trial(i,:) = ...
        NeuralPopResponse(t,A,B,w0,VehicleLatency + latN_RVa,VehicleGain);
    %present
    RVp_trial(i,:) = ...
        NeuralPopResponse(t,A,B,w0,VehicleLatency + latN_RVp,VehicleGain);
    %absent
    RFa_trial(i,:) = ...
        NeuralPopResponse(t,A,B,w0,FaceLatency + latN_RFa,FaceGain);

    %%% Modification of response by the abscence/presence
    RFp_trial(i,:) = RFp_trial(i,:)*preMod;
    RVp_trial(i,:) = RVp_trial(i,:)*preMod;

    %%% Attentional Modulation / Task related
    modulation = AttentionalModulation(t,mini,attentionLatency,maxi,tmax);
    RFpA_trial(i,:) = RFp_trial(i,:) .* modulation;
    RFaA_trial(i,:) = RFa_trial(i,:) .* modulation;
    RVpA_trial(i,:) = RVp_trial(i,:) .* modulation;
```

```
RVaA_trial(i,:) = RVa_trial(i,:) .* modulation;

% Local inhibition
[RFpA_trial(i,:) RVa_trial(i,:)] = ...
    LocalInhibition(RFpA_trial(i,:),RVa_trial(i,:),factorInhib);
[RFaA_trial(i,:) RVp_trial(i,:)] = ...
    LocalInhibition(RFaA_trial(i,:),RVp_trial(i,:),factorInhib);
[RVpA_trial(i,:) RFa_trial(i,:)] = ...
    LocalInhibition(RVpA_trial(i,:),RFa_trial(i,:),factorInhib);
[RVaA_trial(i,:) RFp_trial(i,:)] = ...
    LocalInhibition(RVaA_trial(i,:),RFp_trial(i,:),factorInhib);

% Normalization of all values
MAXIMUM = ...
    max([RFpA_trial(i,:) RFaA_trial(i,:)...
     RVpA_trial(i,:) RVaA_trial(i,:)]);
RFpA_trial(i,:)=RFpA_trial(i,:)/MAXIMUM;
RFaA_trial(i,:)=RFaA_trial(i,:)/MAXIMUM;
RVpA_trial(i,:)=RVpA_trial(i,:)/MAXIMUM;
RVaA_trial(i,:)=RVaA_trial(i,:)/MAXIMUM;
RFp_trial(i,:)=RFp_trial(i,:)/MAXIMUM;
RVa_trial(i,:)=RVa_trial(i,:)/MAXIMUM;
RVp_trial(i,:)=RVp_trial(i,:)/MAXIMUM;
RFa_trial(i,:)=RFa_trial(i,:)/MAXIMUM;

%%% Combination
RFTarget_trial(i,:) = RFpA_trial(i,:) + RVa_trial(i,:);
RFDistra_trial(i,:) = RFaA_trial(i,:) + RVp_trial(i,:);
RVTarget_trial(i,:) = RVpA_trial(i,:) + RFa_trial(i,:);
RVDistra_trial(i,:) = RVaA_trial(i,:) + RFp_trial(i,:);

% Normalization of sensory values
MAXIMUM = ...
    max([RFTarget_trial(i,:) RFDistra_trial(i,:)...
     RVTarget_trial(i,:) RVDistra_trial(i,:)]);
RFTarget_trial(i,:) = RFTarget_trial(i,:) / MAXIMUM;
RFDistra_trial(i,:) = RFDistra_trial(i,:) / MAXIMUM;
RVTarget_trial(i,:) = RVTarget_trial(i,:) / MAXIMUM;
RVDistra_trial(i,:) = RVDistra_trial(i,:) / MAXIMUM;

% FACE
% accumulation
RFCTarget_trial(i,:) = LeakyIntegrator(RFTarget_trial(i,:),leak);
RFCDistra_trial(i,:) = LeakyIntegrator(RFDistra_trial(i,:),leak);

% noise on the accumulation rise rate (gaussian ; center=1, sd = Ns)
RFTarget_trial(i,:) = RFTarget_trial(i,:) * random('norm',1,Ns);
RFDistra_trial(i,:) = RFDistra_trial(i,:) * random('norm',1,Ns);


[RTF(i),correctF(i),RFTarget_trialNm(i,:),RFDistra_trialNm(i,:),S0_F(i)] = ...
    Competition(RFCTarget_trial(i,:),RFCDistra_trial(i,:),Nm,St);

% VEHICLE
RVCTarget_trial(i,:) = LeakyIntegrator(RVTarget_trial(i,:),leak);
RVCDistra_trial(i,:) = LeakyIntegrator(RVDistra_trial(i,:),leak);

RVTarget_trial(i,:) = RVTarget_trial(i,:) * random('norm',1,Ns);
```

```
    RVDistra_trial(i,:) = RVDistra_trial(i,:) * random('norm',1,Ns);

    [RTV(i),correctV(i),RVTarget_trialNm(i,:),RVDistra_trialNm(i,:),S0_V(i)] = ...
        Competition(RVCTarget_trial(i,:),RVCDistra_trial(i,:),Nm,St);
end
```

A.2.2 Paramètres

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                                PARAM
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% DIFFERENCE BETWEEN CATEGORIES
% Latency %in monkey IT = 90 for faces (Kiani et al)
FaceLatency=90;
VehicleLatency=120;

% PARAM PRESENCE/ABSCENCE - gain modulation
preMod = 4; % difference of activation when the image is present Vs absent

% LOCAL MUTUAL INHIBITION
factorInhib = 0.3;

% SENSORY NOISE (gaussian)
Ns = 2;

% MOTOR NOISE (uniform)
Nm = 30;

% LATENCY NOISE (uniform)
Nl = 0; % in ms

% PARAM INTEGRATOR
leak = 0.000; %0.008;
St = 30;
```

A.2.3 Fonction : Réponse sensorielle

```
function [PopResponse] = NeuralPopResponse(t,A,B,w0,popLatency,gain)
% PopResponse = NeuralPopResponse([1:500],0.03,0.5,0.017,100,1)
% gain==1 means no specific gain

t=t(1:(end-popLatency));

PopResponse = (A*t+B).*exp(-w0*t)-B;

% trick to control the latency of the pop response
PopResponse = [zeros(1,popLatency) PopResponse];

% addition of a gain
PopResponse = PopResponse*gain;
end
```

A.2.4 Fonction : Modulation attentionnelle

```
function [modulation] = ...
    AttentionalModulation(t,mini,attentionLatency,maxi,tmax)
%
% controls
if attentionLatency < 2; attentionLatency=2; end;
if
    tmax==attentionLatency;
    error('tmax cannot be = to attentionLatency');
end;
if maxi≤mini; error('max cannot be = or < to min'); end;

% 1ere phase plate
modulation(1:attentionLatency-2) = mini;
% 2eme phase
% on monte vers le max a une vitesse
% dependant du temps qu'on a pour le faire
modulation((attentionLatency-1):tmax-1) = ...
    [mini:((maxi-mini)/(tmax-attentionLatency)):maxi];
% 3eme phase plate
modulation(tmax:length(t)) = maxi;

% smooth
modulation = smooth(modulation,100)';
end
```

A.2.5 Fonction : Inhibition entre les populations pour un même image

```
function [R1new,R2new] = LocalInhibition(R1,R2,factorInhib)

for t = 1:length(R1)
    if R2(t) ≠ 0
        R1new(t) = R1(t) - (R2(t)*factorInhib);
    else R1new(t) = R1(t);
    end
    if R1(t) ≠ 0
        R2new(t) = R2(t) - (R1(t)*factorInhib);
    else R2new(t) = R2(t);
    end
    % remove values <0
    if R1new(t)<0
        R1new(t)=0;
    end
    if R2new(t)<0
        R2new(t)=0;
    end
end
end
```

A.2.6 Fonction : Accumulateur à fuite

```
function [cumleakedR] = LeakyIntegrator(R,leak)
% R = Response, vector of activation along time
% leak = value for the leak. Example: 0.002
```

```
cumleakedR(1) = R(1);

for i=2:length(R)
    cumleakedR(i) = cumleakedR(i-1) + R(i) - ((cumleakedR(i-1))*leak);
end
end
```

A.2.7 Fonction : Compétition entre les 2 alternatives

```
function [RT,correct,Rc,Ri,S0] = Competition(Rc,Ri,Nm,threshold)
% Competition for decision between 2 populations
% Rc = Response for correct response
% Ri = Response for incorrect one

S0 = random('unif', -Nm, Nm);

% The motor noise is the same applied to both side.
% For example we hypothesized that if the subject have a tendency to
% saccade on the left on this trial, he will have an equivalent inhibition
% to saccade on the right on that same trial.

Rc = Rc + S0;
Ri = Ri - S0;

RTc = find(Rc > threshold, 1 );
RTi = find(Ri > threshold, 1 );

if isempty(RTc); RTc = 9999; end;
if isempty(RTi); RTi = 9999; end;

if RTc < RTi
    correct = 1;
    RT = RTc;
elseif RTc > RTi
    correct = 0;
    RT = RTi;
% if they arrive at the same time, then we flip a coin to decide who wins
else
    tirage = CoinFlip(1,0.5);
    if tirage==1
        correct = 1;
        RT = RTc;
    else correct = 0;
        RT = RTi;
    end
end

RT = RT+20; % 20 = time to initiate a saccade from decision
end
```

Bibliographie

Alvarez, G. et Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, 19(4):392.

Alvarez, G. a. et Oliva, a. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, 106(18):7345–7350.

Antal, A., Kéri, S., Kovács, G., Janka, Z. et Benedek, G. (2000). Early and late components of visual categorization : an event-related potential study. *Brain research. Cognitive brain research*, 9(1):117–9.

Bacon-Macé, N., Kirchner, H., Fabre-Thorpe, M. et Thorpe, S. J. (2007). Effects of task requirements on rapid natural scene processing : From common sensory encoding to distinct decisional mechanisms. *Journal of Experimental Psychology : Human Perception and Performance*, 33(5):1013–1026.

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M. et Thorpe, S. J. (2005). The time course of visual processing : Backward masking and natural scene categorisation. *Vision Research*, 45(11):1459–1469.

Bailey, A. J., Braeutigam, S., Jousmäki, V. et Swithenby, S. J. (2005). Abnormal activation of face processing systems at early and intermediate latency in individuals with autism spectrum disorder : a magnetoencephalographic study. *The European journal of neuroscience*, 21(9):2575–85.

Bar, M. (2007). The proactive brain : using analogies and associations to generate predictions. *Trends in cognitive sciences*, 11(7):280–9.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., Shadlen, M. N., Latham, P. E. et Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, 60(6):1142–1152.

Bichot, N. P., Schall, J. D. et Thompson, K. G. (1996). Visual feature selectivity in frontal eye fields induced by experience in mature macaques. *Nature*, 381:697–699.

Bindemann, M., Burton, A., Hooge, I., Jenkins, R. et de Haan, E. (2005). Faces retain attention. *Psychonomic Bulletin & Review*, 12(6):1048.

Bindemann, M., Burton, A., Langton, S., Schweinberger, S. et Doherty, M. (2007). The control of attention to faces. *Journal of Vision*, 7:10–15.

Birmingham, E., Bischof, W. F. et Kingstone, A. (2009). Saliency does not ac-

count for fixations to eyes within social scenes. *Vision Research*, 49(24):2992–3000.

Blatt, G. J., Andersen, R. A. et Stoner, G. R. (1990). Visual Receptive Field Organization and Cortico-Cortical Connections of the Lateral Intraparietal Area (Area LIP) in the Macaque. *The Journal of Comparative Neurology*, 299:421–445.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P. et Cohen, J. D. (2006). The physics of optimal decision making : a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700–65.

Bogacz, R., Wagenmakers, E.-j., Forstmann, B. U. et Nieuwenhuis, S. (2009). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, 33(1):10–16.

Braeutigam, S., Bailey, A. J. et Swithenby, S. J. (2001). Task-dependent early latency (30–60 ms) visual processing of human faces and other objects. *Neuroreport*, 12(7):1531–1536.

Braun, J. et Mattia, M. (2010). Attractors and noise : Twin drivers of decisions and multistability. *NeuroImage*, 52(3):740–751.

Brown, S. et Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, 112(1):117–128.

Brown, S. D. et Heathcote, A. (2008). The simplest complete model of choice response time : linear ballistic accumulation. *Cognitive psychology*, 57(3):153–78.

Brown, V., Huey, D. et Findlay, J. (1997). Face detection in peripheral vision : do faces pop out ? *Perception*, 26:1555–1570.

Brysbaert, M., Drieghe, D. et Vitu, F. (2005). *Cognitive Processes in Eye Guidance.*, chapitre Word skipping : Implications for theories of eye movement control in reading, pages 381–393. Oxford University Press, Oxford.

Bullier, J. (2001). Integrated model of visual processing. *Brain research. Brain research reviews*, 36(2-3):96–107.

Bullier, J. (2004). *The Primate Visual System*, chapitre Hierarchies of cortical areas, pages 181–204.

Bullier, J. et Nowak, L. (1995). Parallel versus serial processing : new vistas on the distributed organization of the visual system. *Current Opinion in Neurobiology*, 5(4):497–503.

Buswell, G. T. (1935). *How People Look at Pictures : A Study of The Psychology of Perception in Art*. Chicago.

Campbell, F. et Robson, J. (1968). Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology*, 197(3):551.

Camprodon, J. A., Zohary, E., Pascual-Leone, A. et Brodbeck, V. (2010). Two Phases of V1 Activity for Visual Recognition of Natural Images. *Journal of cognitive neuroscience*, 22(6):1262–9.

Carmi, R. et Itti, L. (2006). Visual causes versus correlates of attentional selection

in dynamic scenes. *Vision Research*, 46(3):4333–4345.

Carpenter, R. (2004). Contrast, Probability, and Saccadic Latency : Evidence for Independence of Detection and Decision. *Current Biology*, 14:1576–1580.

Cerf, M., Frady, E. et Koch, C. (2009). Faces and text attract gaze independent of the task : Experimental data and computer model. *Journal of Vision*, 9(12):1–15.

Crick, F. (1984). Function of the thalamic reticular complex : the searchlight hypothesis. *Proc. Natl. Acad. Sci. USA*, 81:4586–4590.

Crouzet, S. M., Kirchner, H. et Thorpe, S. J. (2010). Fast saccades towards face : Face detection in just 100 ms. *Journal of Vision*, 10(4):1–17.

Dakin, S. et Watt, R. (2009). Biological bar codes in human faces. *Journal of Vision*, 9(4):1–10.

Davenport, J. L. et Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, 15(8):559–64.

David, S. V., Hayden, B. Y. et Gallant, J. L. (2006). Spectral Receptive Field Properties Explain Shape Selectivity in Area V4. *Journal of Neurophysiology*, 96(6): 3492–3505.

David, S. V., Mazer, J. a., Hayden, B. Y. et Gallant, J. L. (2008). Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron*, 59(3):509–21.

Debener, S., Ullsperger, M., Siegel, M. et Engel, A. K. (2006). Single-trial EEG-fMRI reveals the dynamics of cognitive function. *Trends in cognitive sciences*, 10(12):558–63.

Delorme, A., Richard, G. et Fabre-thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues : A study in monkeys and humans.

Desimone, R., Albright, T. D., Gross, C. G. et Bruce, C. (1984). Stimulus-selective properties of Inferior Temporal Neurons in the Macaque. *The Journal of Neuroscience*, 4(8):2051–2062.

Desimone, R. et Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18:193–222.

Desimone, R. et Schein, S. (1987). Visual properties of neurons in area V4 of the macaque : sensitivity to stimulus form. *Journal of neurophysiology*, 57(3):835–868.

DiCarlo, J. J. et Maunsell, J. H. R. (2005). Using neuronal latency to determine sensory-motor processing pathways in reaction time tasks. *Journal of neurophysiology*, 93(5):2974–86.

Drewes, J., Trommershaeuser, J. et Gegenfurtner, K. R. (2009). The effect of context on rapid animal detection. *[Abstract] Journal of Vision*, 9(8):1117.

Duncan, J. et Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological review*, 96(3):433–58.

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A. et Oliva, A. (2009). Modelling

search for people in 900 scenes : A combined source model of eye guidance. *Visual Cognition*, 17(6):945–978.

Einhäuser, W., König, P., Rutishauser, U., Frady, E. P., Nadler, S. et Koch, C. (2006). The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of vision*, 6(11):1148–58.

Einhäuser, W., Rutishauser, U. et Koch, C. (2008a). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 8(2):1–19.

Einhäuser, W., Spain, M. et Perona, P. (2008b). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):1–26.

Evans, K. K. et Treisman, A. (2005). Perception of objects in natural scenes : is it really attention free ? *Journal of experimental psychology. Human perception and performance*, 31(6):1476–92.

Fabre-Thorpe, M., Delorme, A., Marlot, C. et Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of cognitive neuroscience*, 13(2):171–80.

Fei-Fei, L., Iyer, A., Koch, C. et Perona, P. (2007). What do we perceive in a glance of a real-world scene ? *Journal of vision*, 7(1):1–29.

Fei-Fei, L., VanRullen, R., Koch, C. et Perona, P. (2005). Why does natural scene categorization require little attention ? Exploring attentional requirements for natural and synthetic stimuli. *Visual Cognition*, 12(6):893–924.

Felleman, D. J. et Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex*, 1:1–47.

Felsen, G. et Dan, Y. (2005). A natural approach to studying vision. *Nature neuroscience*, 8(12):1643–6.

Ferrera, V. P., Yanike, M. et Cassanello, C. (2009). Frontal eye field neurons signal changes in decision criteria. *Nature neuroscience*, 12(11):1458–62.

Findlay, J. M. et Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *The Behavioral and brain sciences*, 22(4):661–74 ; discussion 674–721.

Fischer, B. et Weber, H. (1993). Express Saccades and Visual Attention. *Behavioral and Brain Sciences*, 16(3):553–567.

Fize, D., Boulanouar, K., Chatel, Y., Ranjeva, J., Fabre-Thorpe, M. et Thorpe, S. (2000). Brain areas involved in rapid categorization of natural images : An event-related fMRI study. *Neuroimage*, 11(6):634–643.

Freedman, D. J. et Assad, J. a. (2006). Experience-dependent representation of visual categories in parietal cortex. *Nature*, 443(7107):85–8.

Fries, W. (1981). The projection from the lateral geniculate nucleus to the prestriate cortex of the macaque monkey. *Proceedings of the Royal Society of London. Se-*

ries B, Containing papers of a Biological character. Royal Society (Great Britain), 213(1190):73–86.

Gallant, J., Connor, C., Rakshit, S., Lewis, J. W. et Van Essen, D. (1996). Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *Journal of neurophysiology*, 76(4):2718–2739.

Gallant, J. L., Connor, C. E. et Essen, D. C. V. (1998). Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. *NeuroReport*, 9(1):1673–1678.

Gallant, J. L., Shoup, R. E. et Mazer, J. A. (2000). A human extrastriate area functionally homologous to macaque V4. *Neuron*, 27(2):227–35.

Garrido, L., Duchaine, B. et Nakayama, K. (2008). Face detection in normal and prosopagnosic individuals. *Journal of Neuropsychology*, 2(1):119–140.

Gaspar, C. M. et Rousselet, G. A. (2009). How do amplitude spectra influence rapid animal detection ? *Vision research*, 49(24):3001–12.

Gautrais, J. et Thorpe, S. (1998). Rate coding versus temporal order coding : a theoretical approach. *Bio Systems*, 48(1-3):57–65.

Gegenfurtner, K. R. et Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current biology : CB*, 10(13):805–8.

Geisler, W. S. (2008). Visual Perception and the Statistical Properties of Natural Scenes. *Annual Review of Psychology*, 59(1):167–192.

Ghose, G. M. et Ts'O, D. Y. (1997). Form Processing Modules in Primate Area V4. *Journal of Neurophysiology*, 77:2191–2196.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Psychology Press, lawrence e édition.

Girard, P., Jouffrais, C. et Kirchner, C. H. (2008). Ultra-rapid categorisation in non-human primates. *Animal cognition*, 11(3):485–93.

Glimcher, P. W. (2001). Making choices : the neurophysiology of visual-saccadic decision making. *Trends in neurosciences*, 24(11):654–9.

Gold, J. et Shadlen, M. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30:535–574.

Gollisch, T. et Meister, M. (2010). Eye Smarter than Scientists Believed : Neural Computations in Circuits of the Retina. *Neuron*, 65(2):150–164.

Graham, D. J. et Field, D. J. (2007). *Efficient Coding of Natural Images*. Academic Press.

Grill-Spector, K. (2003). The neural basis of object perception. *Current Opinion in Neurobiology*, 13(2):159–166.

Grill-Spector, K. et Kanwisher, N. (2005). Visual Recognition. As soon as you know it is there, you know what it it. *Psychological Science*, 16(2):152–160.

Grill-Spector, K., Kourtzi, Z. et Kanwisher, N. (2001). The lateral occipital

complex and its role in object recognition. *Vision research*, 41(10-11):1409–22.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y. et Malach, R. (1998). A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Human brain mapping*, 6(4):316–28.

Grill-Spector, K. et Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, 27:649–77.

Gross, C., Rocha-miranda, C. et Bender, D. (1972). Visual Cortex Properties of Neurons in Inferotemporal of the Macaque. *Journal of Neurophysiology*, 35(1):96–111.

Guyader, N. (2004). Image phase or amplitude ? Rapid scene categorization is an amplitude-based process. *Comptes Rendus Biologies*, 327(4):313–318.

Guyonneau, R., Kirchner, H. et Thorpe, S. (2006). Animals roll around the clock : The rotation invariance of ultrarapid visual processing. *J Vis*, 6(10):1008–17.

Guyonneau, R., VanRullen, R. et Thorpe, S. (2005). Neurons tune to the earliest spikes through STDP. *Neural Computation*, 17(4):859–879.

Hanes, D. et Schall, J. (1996). Neural control of voluntary movement initiation. *Science*, 274(5286):427.

Hanks, T., Ditterich, J. et Shadlen, M. (2006). Microstimulation of macaque area lip affects decision-making in a motion discrimination task. *Nat. Neurosci.*, 9:682–9.

Hansen, K. a., Kay, K. N. et Gallant, J. L. (2007). Topographic organization in and near human visual area V4. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(44):11896–911.

Heekeren, H., Marrett, S. et Ungerleider, L. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, 9(6):467–480.

Henderson, J. et Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology*, 50(1):243–271.

Henderson, J. M. (2007). Regarding Scenes. *Current Directions in Psychological Science*, 16(4):219–222.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S. et Mack, M. (2007). *Visual saliency does not account for eye movements during visual search in real-world scenes*, chapitre 25, pages 537–562. Elsevier Ltd.

Hershler, O. et Hochstein, S. (2005). At first sight : A high-level pop out effect for faces. *Vision Research*, 45(13):1707–1724.

Hershler, O. et Hochstein, S. (2006). With a careful look : Still no low-level confound to face pop-out. *Vision Research*, 46:3028–3035.

Honey, C., Kirchner, H. et VanRullen, R. (2008). Faces in the cloud : Fourier power spectrum biases ultrarapid face detection. *Journal of Vision*, 8(12):9.

Horwitz, G., Batista, A. et Newsome, W. (2004a). Representation of an abstract

perceptual decision in macaque superior colliculus. *J. Neurophysiol.*, 91:2281–96.

Horwitz, G. D., Batista, A. P. et Newsome, W. T. (2004b). Direction-selective visual responses in macaque superior colliculus induced by behavioral training. *Neuroscience letters*, 366(3):315–9.

Hsieh, P.-J., Vul, E. et Kanwisher, N. (2010). Recognition alters the spatial pattern of FMRI activation in early retinotopic cortex. *Journal of neurophysiology*, 103(3): 1501–7.

Hubel, D. H. et Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148:574–591.

Hung, C., Kreiman, G., Poggio, T. et DiCarlo, J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310:863–866.

Hupé, J.-M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P. et Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*, 394(August):784–787.

Itti, L. (2005). Quantifying the contribution of low,Äêlevel saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123.

Itti, L. et Koch, C. (2001). Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3):194–203.

Itti, L., Koch, C., Niebur, E. et Others (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

Janssen, P., Orban, G. a., Srivastava, S. et Ombelet, S. (2008). Coding of shape and position in macaque lateral intraparietal area. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(26):6679–90.

Johnson, J. S. et Olshausen, B. a. (2003). Timecourse of neural signatures of object recognition. *Journal of vision*, 3(7):499–512.

Johnson, M. H. (2005). Subcortical face processing. *Nature reviews. Neuroscience*, 6(10):766–74.

Jolicoeur, P., Gluck, M. a. et Kosslyn, S. M. (1984). Pictures and names : making the connection. *Cognitive psychology*, 16(2):243–75.

Joubert, O., Rousselet, G., Fize, D. et Fabre-Thorpe, M. (2007). Processing scene context : Fast categorization and object interference. *Vision research*, 47(26): 3286–3297.

Joubert, O. R., Fize, D., Rousselet, G. A. et Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision*, 8(13):1–18.

Kamitani, Y. et Tong, F. (2006). Decoding seen and attended motion directions from activity in the human visual cortex. *Current biology*, 16(11):1096–102.

Kanan, C., Tong, M. H., Zhang, L. et Cottrell, G. W. (2009). SUN : Top-down

saliency using natural statistics. *Visual Cognition*, 17:979–1003.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature neuroscience*, 3(8):759–63.

Kanwisher, N., McDermott, J. et Chun, M. (1997). The fusiform face area : a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11):4302.

Kaping, D., Tzvetanov, T. et Treue, S. (2007). Adaptation to statistical properties of visual scenes biases rapid categorization. *Visual Cognition*, 15(1):12–19.

Kayser, C. (2004). Processing of complex stimuli and natural scenes in the visual cortex. *Current Opinion in Neurobiology*, 14(4):468–473.

Kayser, C., Konig, P. et Salazar, R. F. (2003). Responses to natural scenes in cat V1. *Journal of neurophysiology*, 90(3):1910–20.

Kayser, C., Nielsen, K. J. et Logothetis, N. K. (2006). Fixations in natural scenes : interaction of image structure and image content. *Vision research*, 46(16):2535–45.

Keightley, M. L., Winocur, G., Graham, S. J., Mayberg, H. S., Hevenor, S. J. et Grady, C. L. (2003). An fMRI study investigating cognitive modulation of brain regions associated with emotional processing of visual stimuli. *Neuropsychologia*, 41(5):585–96.

Keil, M. (2008). Does face image statistics predict a preferred spatial frequency for human face processing ? *Proceedings of the Royal Society B : Biological Sciences*, 275(1647):2095.

Kiani, R., Esteky, H., Mirpour, K. et Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of neurophysiology*, 97(6):4296–309.

Kiani, R., Esteky, H. et Tanaka, K. (2005). Differences in Onset Latency of Macaque Inferotemporal Neural Responses to Primate and Non-Primate Faces. *Journal of Neurophysiology*, 94:1587–1596.

Kiani, R., Hanks, T. D. et Shadlen, M. N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(12):3017–29.

Kim, J. N. et Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nat. Neurosci.*, 2:176–185.

Kirchner, H., Barbeau, E. J., Thorpe, S. J., Régis, J. et Liégeois-Chauvel, C. (2009). Ultra-rapid sensory responses in the human frontal eye field region. *The Journal of Neuroscience*, 29(23):7599–606.

Kirchner, H. et Thorpe, S. (2006). Ultra-rapid object detection with saccadic eye movements : Visual processing speed revisited. *Vision Research*, 46(11):1762–1776.

Koch, C. (2004). *The Quest For Consciousness, a neurobiological approach*. Roberts

& Company.

Kreiman, G., Fried, I. et Koch, C. (2002). Single-neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):8378–83.

Kreiman, G., Koch, C. et Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, 3(9):946–953.

Lamme, V. et Roelfsema, P. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579.

Lee, T. et Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448.

Lee, T. S., Yang, C. F., Romero, R. D. et Mumford, D. (2002). Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. *Nature Neuroscience*, 5(6):589–597.

Lesica, N. a. et Stanley, G. B. (2004). Encoding of natural scene movies by tonic and burst spikes in the lateral geniculate nucleus. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 24(47):10731–40.

Levi, D. M. (2008). Crowding–an essential bottleneck for object recognition : a mini-review. *Vision research*, 48(5):635–54.

Lewis, M. et Edmonds, A. (2005). Searching for faces in scrambled scenes. *Visual Cognition*, 12(7):1309–1336.

Lewis, M. et Ellis, H. (2003). How we detect a face : A survey of psychological evidence. *International Journal of Imaging Systems and Technology*, 13(1):3–7.

Lewis, M. B. et Edmonds, A. J. (2003). Face detection : Mapping human performance. *Perception*, 32(8):903–920.

Li, F., VanRullen, R., Koch, C. et Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99(14):9596.

Liu, H., Agam, Y., Madsen, J. R. et Kreiman, G. (2009). Timing, Timing, Timing : Fast Decoding of Object Information from Intracranial Field Potentials in Human Visual Cortex. *Neuron*, 62:281–290.

Lo, C.-C. et Wang, X.-J. (2006). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature neuroscience*, 9(7):956–63.

Logothetis, N. et Sheinberg, D. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19(1):577–621.

Lyon, D. C., Nassi, J. J. et Callaway, E. M. (2010). A Disynaptic Relay from Superior Colliculus to Dorsal Stream Visual Cortex in Macaque Monkey. *Neuron*, 65(2):270–279.

Macé, M. J.-M., Joubert, O. R., Nespoulous, J.-L. et Fabre-Thorpe, M. (2009).

The time-course of visual categorizations : you spot the animal faster than the bird. *PloS one*, 4(6):e5927.

Macé, M. J.-M., Thorpe, S. J. et Fabre-Thorpe, M. (2005). Rapid categorization of achromatic natural scenes : how robust at very low contrasts ? *The European journal of neuroscience*, 21(7):2007–18.

Mack, M. L., Gauthier, I., Sadr, J. et Palmeri, T. J. (2008). Object detection and basic-level categorization : Sometimes you know it is there before you know what it is. *Psychonomic Bulletin & Review*, 15(1):28–35.

Mack, M. L., Wong, A. C.-N., Gauthier, I., Tanaka, J. W. et Palmeri, T. J. (2009). Time course of visual object categorization : fastest does not necessarily mean first. *Vision research*, 49(15):1961–8.

Malcolm, G. L. et Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes : Evidence from eye movements. *Journal of Vision*, 9(11):1–13.

Mante, V., Bonin, V. et Carandini, M. (2008). Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron*, 58(4):625–38.

Marr, D. (1982). *Vision : A computational investigation into the human representation and processing of visual information*. W.H.Freeman & Co Ltd, San Francisco.

Masquelier, T., Guyonneau, R. et Thorpe, S. J. (2009). Competitive STDP-based spike pattern learning. *Neural computation*, 21(5):1259–76.

Masquelier, T. et Thorpe, S. J. (2007). Unsupervised Learning of Visual Features through Spike Timing Dependent Plasticity. *PLoS Computational Biology*, 3(2):e31.

Maunsell, J. H. et Treue, S. (2006). Feature-based attention in visual cortex. *TRENDS in Neurosciences*, 29(6):317–322.

Mazer, J. a. et Gallant, J. L. (2003). Goal-related activity in V4 during free viewing visual search. Evidence for a ventral stream visual salience map. *Neuron*, 40(6):1241–50.

Mazurek, M. E., Roitman, J. D., Ditterich, J. et Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cerebral Cortex*, 13:1257–1269.

McAlonan, K., Cavanaugh, J. et Wurtz, R. H. (2008). Guarding the gateway to cortex with attention in visual thalamus. *Nature*, 456:391–394.

Mervis, C. B. et Rosch, E. (1981). Categorization of Natural Objects. *Annual Review of Psychology*, 32(1):89–115.

Mishkin, M., Ungerleider, L. G. et Macko, K. A. (1983). Object vision and spatial vision : two cortical pathways. *Trends in Neurosciences*, 6:414–417.

Mohler, C. W., Goldberg, M. E. et Wurtz, R. H. (1973). Visual receptive fields of frontal eye field neurons. *Brain research*, 61:385–9.

Monosov, I. E., Trageser, J. C. et Thompson, K. G. (2008). Measurements of simultaneously recorded spiking activity and local field potentials suggest that spatial

selection emerges in the frontal eye field. *Neuron*, 57(4):614–25.

Moss, H. E., Rodd, J. M., Stamatakis, E. A., Bright, P. et Tyler, L. K. (2005). Anteromedial temporal cortex supports fine-grained differentiation among objects. *Cerebral cortex*, 15(5):616–27.

Mumford, D. (1991). On the computational architecture of the neocortex : I. The role of thalamo-cortical loop. *Biological Cybernetics*, 65:135–145.

Mumford, D. (1992). On the computational architecture of the neocortex : II The role of cortico-cortical loops. *Biological Cybernetics*, 66:241–251.

Mur, M., Bandettini, P. et Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI–an introductory guide. *Social Cognitive and Affective Neuroscience*, pages 1–22.

Nandakumar, C. et Malik, J. (2009). Understanding rapid category detection via multiply degraded images. *Journal of Vision*, 9(6):1–8.

Nestor, A., Vettel, J. M. et Tarr, M. J. (2008). Task-Specific Codes for Face Recognition : How they Shape the Neural Representation of Features for Detection and Individuation. *PloS one*, 3(12).

New, J., Cosmides, L. et Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, 104(42):16598.

Norman, K., Polyn, S., Detre, G. et Haxby, J. (2006). Beyond mind-reading : multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430.

Nothdurft, H. (1993). Faces and facial expressions do not pop out. *Perception*, 22:1287–1298.

Oliva, A. et Torralba, A. (2001). Modeling the shape of the scene : A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Oliva, A. et Torralba, A. (2006). Building the gist of a scene : The role of global image features in recognition. *Progress in Brain Research*, 155(1):23.

Oppenheim, A. et Lim, J. (1981). The Importance of Phase in Signals. *Proceedings of the IEEE*, 69(5):529–541.

Parkhurst, D., Law, K. et Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123.

Pascalis, O. (1998). Face recognition in primates : a cross-species study. *Behavioural Processes*, 43(1):87–96.

Pascalis, O., de Haan, M. et Nelson, C. a. (2002). Is face processing species-specific during the first year of life ? *Science (New York, N.Y.)*, 296(5571):1321–3.

Peelen, M., Fei-Fei, L. et Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, 460(7251):94–97.

Pelli, D. (2008). Crowding : a cortical constraint on object recognition. *Current Opinion in Neurobiology*, 18(4):445–451.

Pelli, D. et Tillman, K. (2008). The uncrowded window of object recognition. *Nature neuroscience*, 11(10):1129–1135.

Peng, X., Sereno, M. E., Silva, A. K., Lehky, S. R. et Sereno, A. B. (2008). Shape selectivity in primate frontal eye field. *Journal of Neurophysiology*, 100(2):796–814.

Pereira, F., Mitchell, T. et Botvinick, M. (2009). Machine learning classifiers and fMRI : a tutorial overview. *Neuroimage*, 45(1S1):199–209.

Perret, D., Rolls, E. T. et Caan, W. (1982). Visual Neurones Responsive to Faces in the Monkey Temporal Cortex. *Experimental Brain Research*, 47:329–342.

Perrett, D. I., Oram, M. W. et Wachsmuth, E. (1998). Evidence accumulation in cell populations responsive to faces : an account of generalisation of recognition without mental transformations. *Cognition*, 67:111–145.

Pinto, N., Cox, D. et DiCarlo, J. (2008). Why is real-world visual object recognition hard. *PLoS Computational Biology*, 4(1).

Piotrowski, L. et Campbell, F. (1982). A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346.

Potter, M. C. et Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, 81(1):10–15.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D. et Palmeri, T. J. (2010). Neurally constrained modeling of decisions.

Quiroga, R. Q., Kreiman, G., Koch, C. et Fried, I. (2008). Sparse but not 'grandmother-cell' coding in the medial temporal lobe. *Trends in cognitive sciences*, 12(3):87–91.

Quiroga, R. Q., Reddy, L., Koch, C. et Fried, I. (2007). Decoding visual inputs from multiple neurons in the human temporal lobe. *Journal of neurophysiology*, 98(4):1997–2007.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. et Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–7.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M. et Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision research*, 42(11):1447–63.

Ratcliff, R. (1978). A Theory of Memory Retrieval. *Psychological Review*, 85(2).

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, 86(3):446–461.

Ratcliff, R., Cherian, A. et Segraves, M. (2003). A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. *Journal of neurophysiology*, 90(3):1392–407.

Ratcliff, R., Hasegawa, Y. T., Hasegawa, R. P., Smith, P. L. et Segraves,

M. A. (2007). Dual Diffusion Model for Single-Cell Recording Data From the Superior Colliculus in a Brightness-Discrimination Task. *Journal of Neurophysiology*, 97:1756–1774.

Ratcliff, R. et Rouder, J. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, 9(5):347356.

Ratcliff, R. et Tuerlinckx, F. (2002). Estimating parameters of the diffusion model : approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic bulletin & review*, 9(3):438–81.

Reddi, B., Asrress, K. et Carpenter, R. (2003). Accuracy, Information, and Response Time in a Saccadic Decision Task. *Journal of Neurophysiology*, 90(5):3538.

Rees, G. (2009). Visual Attention : The Thalamus at the Centre ? *Current biology*, 19(5):R213–214.

Reinagel, P., Godwin, D., Sherman, Murray, S. et Koch, C. (1999). Encoding of Visual Information by LGN Bursts. *Journal of Neurophysiology*, 81:2558–2569.

Riesenhuber, M. et Poggio, T. (1999). Hierarchical models of object recognition in cortex. *nature neuroscience*, 2:1019–1025.

Roelfsema, P., Tolboom, M. et Khayat, P. (2007). Different processing phases for features, figures, and selective attention in the primary visual cortex. *Neuron*, 56(5):785–792.

Roitman, J. et Shadlen, M. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.*, 22: 9475–89.

Rossion, B. et Gauthier, I. (2002). How does the brain process upright and inverted faces ? *Behavioral and cognitive neuroscience reviews*, 1(1):63–75.

Rossion, B., Gauthier, I., Tarr, M. J., Despland, P., Bruyer, R., Linotte, S. et Crommelinck, M. (2000). The N170 occipito-temporal component is delayed and enhanced to inverted faces but not to inverted objects : an electrophysiological account of face-specific processes in the human brain. *Neuroreport*, 11(1):69–74.

Rousselet, G., Fabre-Thorpe, M. et Thorpe, S. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, 5(7):629–630.

Rousselet, G., Macé, M. et Fabre-Thorpe, M. (2003a). Is it an animal ? Is it a human face ? Fast processing in upright and inverted natural scenes. *Journal of Vision*, 3(6):440–455.

Rousselet, G., Thorpe, S. et Fabre-Thorpe, M. (2004a). How parallel is visual processing in the ventral pathway ? *Trends in Cognitive Sciences*, 8(8):363–370.

Rousselet, G., Thorpe, S. et Fabre-Thorpe, M. (2004b). Processing of one, two or four natural scenes in humans : the limits of parallelism. *Vision Research*, 44(9):877–894.

Rousselet, G. A., Thorpe, S. J. et Fabre-Thorpe, M. (2003b). Taking the MAX

from neuronal responses. *Trends in Cognitive Sciences*, 7(3):99–102.

Rust, N. C. et Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, 8(12):1647–1650.

Sabatinelli, D., Lang, P. J., Bradley, M. M., Costa, V. D. et Keil, A. (2009). The timing of emotional discrimination in human amygdala and ventral visual cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(47):14864–8.

Sanders, M., Warrington, E. K., Marshall, J. et Wieskrantz, L. (1974). "Blindsight" : Vision in a field defect. *The Lancet*, 303(7860):707–708.

Schall, J. (2002). The neural selection and control of saccades by the frontal eye field. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 357(1424): 1073.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience*, 2:33–42.

Schiller, P. et Lee, K. (1991). The role of the primate extrastriate area V4 in vision. *Science*, 251(4998):1251–1253.

Sereno, a. B. et Maunsell, J. H. (1998). Shape selectivity in primate lateral intraparietal cortex. *Nature*, 395(6701):500–3.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U. et Poggio, T. (2007a). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, 165:33.

Serre, T., Oliva, A. et Poggio, T. (2007b). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6424–9.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. et Poggio, T. (2007c). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411.

Sewards, T. V. et Sewards, M. a. (2002). Innate visual object recognition in vertebrates : some proposed pathways and mechanisms. *Comparative biochemistry and physiology. Part A, Molecular & integrative physiology*, 132(4):861–91.

Shadlen, M. N., Britten, K. H., Newsome, W. T. et Movshon, J. A. (1996). A Computational Analysis of the Relationship between Neuronal and Behavioral Responses to Visual Motion. *The Journal of Neuroscience*, 16(4):1486–1510.

Shadlen, M. N. et Newsome, W. T. (1996). Motion perception : Seeing and deciding. *Proceedings of the National Academy of Sciences of the United States of America*, 93(January):628–633.

Shadlen, M. N. et Newsome, W. T. (2001). Neural Basis of a Perceptual Decision in the Parietal Cortex (Area LIP) of the Rhesus Monkey. *Journal of Neurophysiology*, 86:1916–1936.

Simoncelli, E. P. et Olshausen, B. A. (2001). Natural Image Statistics And Neural Representation. *Annual Review of Neuroscience*, 24:1193–1216.

Sinha, P., Balas, B. et Ostrovsky, Y. (2007). Discovering faces in infancy. *[Abstract] Journal of Vision*, 7(9):569.

Squire, L. R., Stark, C. E. L. et Clark, R. E. (2004). The medial temporal lobe. *Annual review of neuroscience*, 27:279–306.

Stanford, T. R., Shankar, S., Massoglia, D. P., Costello, M. G. et Salinas, E. (2010). Perceptual decision making in less than 30 milliseconds. *Nature neuroscience*, 13(3):379–386.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19(1):109–139.

Tatler, B. W., Baddeley, R. J. et Gilchrist, I. D. (2005). Visual correlates of fixation selection : effects of scale and time. *Vision research*, 45(5):643–59.

Thorpe, S. et Fabre-Thorpe, M. (2001). Seeking Categories in the Brain. *Science*, 291:260–262.

Thorpe, S., Fize, D. et Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520–522.

Thorpe, S. J. (1990). *Spike arrival times : A highly efficient coding scheme for neural networks*, pages 91–94. North-Holland Elsevier.

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M. et Bülthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. *The European journal of neuroscience*, 14(5):869–76.

Torralba, A. et Oliva, A. (2003). Statistics of natural image categories. *Network : computation in neural systems*, 14(3):391–412.

Torralba, A., Oliva, A., Castelhano, M. S. et Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes : The role of global features in object search. *Psychological Review*, 113(4):766–786.

Treisman, A. et Gelade, G. (1980). A Feature-Integration Theory of Attention. *Cognitive psychology*, 136:97–136.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends in Neurosciences*, 24(5):295–300.

Tsao, D. et Livingstone, M. (2008). Mechanisms of face perception. *Annual review of neuroscience*, 31:411–437.

Tsao, D. Y., Freiwald, W. a., Knutsen, T. a., Mandeville, J. B. et Tootell, R. B. H. (2003). Faces and objects in macaque cerebral cortex. *Nature neuroscience*, 6(9):989–95.

Tyler, L. K., Stamatakis, E. A., Bright, P., Acres, K., Abdallah, S., Rodd, J. M. et Moss, H. E. (2004). Processing objects at different levels of specificity. *Journal of cognitive neuroscience*, 16(3):351–62.

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends in Cognitive Sciences*, 11(2):58–64.

Ullman, S., Vidal-Naquet, M. et Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *nature neuroscience*, 5(7):682–687.

Usher, M. et McClelland, J. L. (2001). The Time Course of Perceptual Choice : The Leaky, Competing Accumulator Model. *Psychological Review*, 108(3):550–592.

VanRullen, R. (2003). Visual saliency and spike timing in the ventral visual pathway. *Journal of Physiology-Paris*, 97(2-3):365–377.

VanRullen, R. (2006). On second glance : Still no high-level pop-out effect for faces. *Vision Research*, 46(18):3017–3027.

VanRullen, R. (2007). The power of the feed-forward sweep. *Advances in Cognitive Psychology*, 3(1-2):167–176.

VanRullen, R. (2009). Binding hardwired versus on-demand feature conjunctions. *Visual Cognition*, 17(1):103–119.

Vanrullen, R., Guyonneau, R. et Thorpe, S. J. (2005). Spike times make sense. *Trends in Neurosciences*, 28(1):1–4.

VanRullen, R., Reddy, L. et Koch, C. (2004). Visual search and dual tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience*, 16(1):4–14.

VanRullen, R. et Thorpe, S. (2001). Is it a bird ? Is it a plane ? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, 30:655–668.

VanRullen, R. et Thorpe, S. (2002). Surfing a spike wave down the ventral stream. *Vision Research*, 42(23):2593–2615.

Vinje, W. et Gallant, J. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273.

Viola, P. et Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Walther, D. et Fei-Fei, L. (2007). Task-set switching with natural scenes : Measuring the cost of deploying top-down attention. *Journal of Vision*, 7(11):1 :12.

Waydo, S., Kraskov, A., Quian Quiroga, R., Fried, I. et Koch, C. (2006). Sparse representation in the human medial temporal lobe. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26(40):10232–4.

Webster, M. J., Bachevalier, J. et Ungerleider, L. G. (1994). Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. *Cerebral Cortex*, 4(5):470–83.

Westheimer, G. (2001). The Fourier theory of vision. *Perception*, 30(1754):531–541.

Wichmann, F. A., Braun, D. I. et Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research*, 46:1520–1529.

Wichmann, F. A., Drewes, J., Rosas, P. et Gegenfurtner, K. R. (2010). Animal detection in natural scenes : Critical features revisited. *Journal of Vision*, 10(4):1–27.

Williams, M. A., Baker, C. I., Beeck, H. P. O. D., Mok, S. W., Dang, S., Triantafyllou, C. et Kanwisher, N. (2008). Feedback of visual object information to foveal retinotopic cortex. *Nature Neuroscience*, 11(12):1439–1445.

Wolfe, J. et Horowitz, T. (2004). What attributes guide the deployment of visual attention and how do they do it ? *Nature Reviews Neuroscience*, 5(6):495–501.

Yang, M.-H. (2009). *Face Detection*, pages 303–308. Li, S. Z.

Yang, M.-H., Kriegman, D. et Ahuja, N. (2002). Detecting faces in images : a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58.

Yantis, S. et Jonides, J. (1984). Abrupt Visual Onsets and Selective Attention : Evidence From Visual Search. *Journal of Experimental Psychology. Human perception and performance*, 10(5):601–621.

Yarbus, A. L. (1967). *Eye Movements and Vision*. Plenum Press, New York.

Zelinsky, G. J. (2008). A Theory of Eye Movements during Target Acquisition. *Psychological Review*, 115(4):787–835.

A QUICK GLANCE AT AN EARLY PHASE OF VISUAL PROCESSING

Author : Sébastien Crouzet **Supervisor :** Dr. Simon J Thorpe

The aim of this thesis is to investigate the dynamics of the cognitive processing involved in rapid object recognition in natural scenes. In order to get the fastest behavioral responses, we used a saccadic choice task in which subjects had to initiate saccades as fast as possible toward the image containing the target among two images displayed at the same time on the screen. This protocol first revealed differences in processing times between categories, with an advantage for the detection of human faces. Indeed, when human faces were used as the target, the first selective saccades appeared as early as 100 ms after the apparition of the images! We were thus interested in the mechanisms allowing such fast detection and showed that a low-level attribute might be used to detect and locate faces in the visual field. In order to understand the nature of the early representation used, we designed two other studies which showed that the fastest saccades were not influenced by contextual information, and were based on relatively coarse information. Finally, I present a simple decision model, based on a latency difference between neuronal population, which accounts for our experimental results. These results, taken in the perspective of what is known about the neural basis of object recognition, showed that the saccadic choice task, allowing access to an early temporal window, will be a very useful tool of interest for future studies on rapid object recognition.

Keywords :
visual perception, natural scenes, ultra-rapid object recognition, saccades, faces, categories

JETER UN REGARD SUR UNE PHASE PRÉCOCE DES TRAITEMENTS VISUELS

Auteur : Sébastien Crouzet **Directeur de thèse :** Dr. Simon J Thorpe

L'objectif de cette thèse a été d'étudier la dynamique des traitements cognitifs permettant la reconnaissance rapide d'objets dans les scènes naturelles. Afin d'obtenir des réponses comportementales précoces, nous avons utilisé un protocole de choix saccadique, dans lequel les sujets devaient diriger leur regard le plus rapidement possible vers l'image contenant l'objet cible parmi deux images affichées à l'écran. Ce protocole a d'abord permis de mettre en évidence des différences de temps de traitement entre les catégories d'objets, avec un avantage particulier pour la détection des visages humains. En effet, lorsque ceux-ci sont utilisés comme cible, les premières saccades sélectives apparaissent dès 100 ms ! Nous nous sommes donc intéressés aux mécanismes permettant une détection aussi rapide et avons montré qu'un attribut bas-niveau pourrait être utilisé pour détecter et localiser les visages dans notre champ visuel en une fraction de seconde. Afin de mieux comprendre la nature des représentations précoces mises en jeu, nous avons mené deux nouvelles études qui nous ont permis de montrer que les saccades les plus rapides ne seraient pas influencées par les informations contextuelles, et seraient basées sur une information rudimentaire. Enfin, j'ai proposé un modèle simple de décision, basé sur des différences de temps de traitement neuronal entre catégories, qui permet de reproduire fidèlement nos résultats expérimentaux. L'ensemble de ces résultats, mis en perspective avec les connaissances actuelles sur les bases neuronales de la reconnaissance d'objet, démontre que le protocole de choix saccadique, en donnant accès à une fenêtre temporelle inaccessible jusqu'alors par les études comportementales, s'avère un outil de choix pour les recherches à venir sur la reconnaissance rapide d'objets.

Mots-clés :
perception visuelle, scènes naturelles, reconnaissance ultra-rapide d'objets, saccades, visages, catégories

Discipline : Neurosciences
Date et lieu : Le 12 juillet 2010 à l'Université Paul Sabatier Toulouse III
Laboratoire : Centre de Recherche Cerveau et Cognition, UMR 5549 (CNRS-Université Paul Sabatier Toulouse 3), Faculté de Médecine de Rangueil, 31062 Toulouse Cedex 9