Cite as : Tarroux, P., & Auvray, M. (2012). La perception visuelle. In M. Denis (Ed.), *La Psychologie Cognitive* (pp. 39-69), Paris : Editions de la Maison des Sciences de l'Homme.

La perception visuelle

Philippe Tarroux[1, 2], Malika Auvray[1]

1. LIMSI-CNRS - Cognition, Perception, Usages - BP 133 91470 Orsay Cedex - France
2. ENS 45 rue d'Ulm 75230 Paris Cedex 05

Résumé

Le processus de perception est celui par lequel tout organisme interagit avec son environnement. Il s'inscrit dans un contexte d'évolution qui en détermine la complexité, du simple filtrage de l'information à la construction de représentations cognitives élaborées. Au travers des capacités attentionnelles, la perception s'appuie sur l'action pour extraire des informations sur le monde environnant. Par ailleurs, la modulation de l'information sensorielle par le contexte d'interaction de l'observateur avec le monde et son enrichissement par les informations préalables dont il dispose font de la perception un processus situé.

Ces mécanismes font actuellement l'objet de tentatives de modélisation et de descriptions formelles qui trouvent un écho dans le développement de systèmes artificiels de perception. Suivant le niveau de réalisme biologique adressé, ces modèles se déclinent suivant divers paradigmes : théorie du filtrage, modèles bayésiens ou encore modèles neuronaux. Ces modèles offrent autant de façons d'appréhender les processus perceptifs, d'en cerner la complexité et de tenter d'en reproduire les propriétés essentielles. L'objectif de ce chapitre est de décrire, sans prétendre à l'exhaustivité, les principaux processus de la perception visuelle en mêlant trois approches : théorique, empirique et appliquée (notamment dans le domaine de la robotique). Seront en particulier abordés les mécanismes de bas niveau, les modèles de reconnaissance d'objets, l'attention visuelle, les liens entre la perception et l'action et le caractère multisensoriel des processus perceptifs.

1. Introduction

Un élément essentiel bien établi désormais est qu'il n'existe pas de problème général de la perception, comme il n'existe pas de problème général de la cognition. Cette assertion fait écho à l'abandon, par l'intelligence artificielle, de la recherche d'un système général de résolution de problèmes. Réduire les processus cognitifs à un unique algorithme est vain. Les systèmes biologiques ont résolu, durant l'évolution, de multiples problèmes dans lesquels cognition et perception jouaient des rôles plus ou moins importants, mais à chaque fois à l'aide de dispositions spécifiques. L'anatomie comparée de l'œil camérulaire, apparu plusieurs fois au cours de l'évolution, l'organisation de la fovéa adaptée aux modes de vie et aux comportements des diverses espèces ou les fluctuations évolutives de la vision des

couleurs en sont des exemples évidents. De même, chez les batraciens, dépourvus de néocortex, les fonctions perceptives essentielles, telles que la capacité à distinguer une proie et un partenaire sexuel, sont implémentées de façon efficace dans les circuits de la rétine et du tectum (Ewert & Arbib, 1989). Il n'en demeure pas moins que ces animaux sont parfaitement adaptés à leur milieu.

Les mécanismes perceptifs se comprennent donc en lien avec l'évolution de l'environnement dans lequel l'organisme se situe. Cependant, dans un environnement donné, deux approches peuvent être considérées. Selon une conception mentaliste, la perception est un processus dans lequel le système sensoriel reçoit passivement des stimulations, puis traite ces informations afin d'identifier objets et évènements sous la forme d'une représentation interne. Le cognitivisme classique demeure globalement dans la lignée de cette conception. La cognition est alors identifiée à un système de traitement linéaire de l'information où les données des sens sont considérées comme une entrée ; à cette entrée succèdent la perception et le raisonnement, qui débouchent sur l'action, envisagée comme une sortie. Née avec la théorie fonctionnaliste du raisonnement comme manipulation de représentations symboliques, cette conception s'appuie sur une cognition envisagée comme un calcul sur des représentations mentales. La tradition mentaliste met ainsi en avant le raisonnement logique comme l'élément essentiel permettant d'accéder au sens. Elle est cependant en échec lorsqu'il s'agit d'expliquer et a fortiori d'implémenter la façon dont les symboles qu'elle manipule sont ancrés dans le réel (Harnad, 1990).

Selon une conception active ou écologique, la perception est construite sur la base de l'extraction de régularités entre les actions effectuées et les stimulations sensorielles résultantes. Parce qu'elle inclut les mécanismes qui permettent d'appréhender le monde extérieur et de construire les modèles d'interaction avec lui (Agre, 1988), la perception participe de façon centrale au processus cognitif. Elle réunit un ensemble de processus qui, dans les diverses modalités sensorielles puis dans leurs combinaisons, conduisent à notre connaissance du monde et à nos aptitudes à nous y comporter de façon cohérente. Ainsi, cette approche qui fait porter à la perception l'essentiel de la mise en cohérence de notre fonctionnement interne avec les caractéristiques du monde extérieur, met l'accent sur le rôle de cette interaction avec le monde comme un élément essentiel de nos aptitudes cognitives.

Notre conception des mécanismes perceptifs est donc influencée par notre position vis-à-vis de la nature essentielle des processus cognitifs. Selon que l'on privilégie une vision mentaliste ou une vision écologique du fonctionnement cognitif, on sera conduit à donner aux processus perceptifs une importance plus ou moins grande dans le processus de construction et d'appropriation du sens. L'objectif de ce chapitre est de présenter une partie des mécanismes qui sous-tendent les processus perceptifs et les modèles qui ont été envisagés suivant l'une ou l'autre des conceptions décrites ci-dessus. Nous abordons ainsi successivement les mécanismes de bas niveau, les modèles de reconnaissance d'objets et de scènes, l'attention visuelle, les liens entre la perception et l'action et les liens entre la vision et les autres modalités sensorielles. Nous verrons que la distinction entre les différentes conceptions de la perception s'illustre par la prééminence de la tradition mentaliste s'agissant de la description des processus de bas niveaux et une plus grande présence de théories alternatives lorsqu'il s'agit de rendre compte des relations entre perception et action comme de tout mécanisme faisant intervenir des processus incarnés et contextualisés.

2. Le filtrage de bas niveau

2.1. Modèles empiriques

Le système visuel est classiquement conçu comme projetant le flux informationnel sur une voie de traitement ventrale aboutissant au cortex inféro-temporal et une voie dorsale conduisant au cortex pariétal (Ungerleider & Mishkin, 1982). Très schématiquement, on associe la voie temporale plutôt à la forme et la voie pariétale à tout ce qui concerne la position. Notons également, parmi les aires visuelles, les champs oculaires frontaux en relation avec la commande des mouvements oculaires. Au-delà des aires visuelles proprement dites, de nombreuses aires associent les informations visuelles à d'autres types de traitements. La conception classique selon laquelle le flux informatif chemine de la rétine vers le cortex est cependant remise en cause par l'abondance des connexions récurrentes qui ont fait progressivement abandonner un schéma de connexions orientées de manière unidirectionelle au profit d'une organisation bidirectionnelle (Van Essen, Anderson, & Felleman, 1992).

A la suite des travaux d'Hübel et Wiesel (1962), on a conçu le système visuel comme un ensemble de filtres destinés à extraire du flux d'information des structures de plus en plus complexes construites par assemblage de structures de plus bas niveau. Cette conception occulte une grande partie du processus perceptif, en particulier dans sa relation à l'espace et dans sa relation à l'action. Toutefois cette théorie du filtrage de l'information visuelle joue un rôle important dans les premières étapes de traitement et a fait l'objet ces dernières années d'importantes avancées. C'est le premier point auquel nous allons nous intéresser. Nous nous contenterons cependant de résumer schématiquement ces résultats en observant que les codages mis en œuvre dans la rétine et dans le cortex V1 conduisent à séparer luminance et chrominance, à coder les oppositions de couleur et à éclater l'information visuelle selon les modalités de luminance, d'orientation et de couleur dans une gamme assez large de fréquences spatiales.

Les observations expérimentales montrent que la rétine réalise plusieurs fonctions essentielles : un échantillonnage du signal, une séparation du codage de la couleur et de la luminance et un filtrage de type rehaussement de contours. Pour le traitement de la couleur, on notera que la présence de deux ou trois types de photorécepteurs rétiniens sensibles à des longueurs d'ondes différentes ne permet pas à elle seule de discriminer les couleurs, un photorécepteur répondant de la même façon à un stimulus faiblement lumineux mais optimal pour sa longueur d'onde et à un stimulus sous-optimal mais fortement lumineux. Pour coder les couleurs, une comparaison entre les réponses d'au moins deux types de ces cellules est nécessaire, ce que réalisent les cellules sensibles à des oppositions de couleur.

Dans le cortex visuel primaire, le schéma instauré par Hübel et Wiesel (1962) a aboutit à distinguer des cellules dites « simples » sensibles à un stimulus orienté et à sa phase par rapport au champ récepteur de la cellule, des cellules « complexes » peu sensibles à la phase et des cellules « hypercomplexes » sensibles aux terminaisons et aux courbures. Adelson et Bergen (1985) ont proposé que les cellules complexes calculent la norme moyenne des sorties des cellules simples (modèle d'énergie) dans une région donnée, ce qui explique leur indépendance en phase et la taille plus élevée de leurs champs récepteurs par rapport aux cellules simples. En dépit de cette caractérisation fonctionnelle, ces distinctions n'ont cependant pas de corrélats anatomiques évidents.

Au-delà de ces aires, les informations de couleur, de disparité binoculaire, d'orientation et de luminance sont recombinées dans V2 et V4 puis dans le cortex inférotemporal conduisant à

des détecteurs de formes complexes. Dans la voie pariétale, le cortex V3 participe au codage du mouvement dont de nombreux modèles ont été proposés (Adelson & Bergen, 1985 ; Heeger, Simoncelli, & Movshon, 1996). Le cortex pariétal fait également l'objet de nombreux travaux montrant par exemple que certaines aires pariétales combinent des informations visuelles, attentionnelles et motrices (Avillac, Denève, Olivier, Pouget, & Duhamel, 2005). Une première approche de la vision consiste donc à retenir, dans une conception ascendante, seulement les aspects de traitement du signal visuel qui conduiraient du signal au sens. Ainsi interprétée, la vision apparaît bien comme un processus de filtrage destiné à extraire du flux visuel une représentation complète de la scène observée codée de la façon la plus optimale possible.

Parmi les modèles formels proposés pour mieux cerner les éléments essentiels de ce processus de codage ascendant, on retiendra ceux qui visent à rendre compte de la sélectivité à l'orientation des cellules du corps genouillé latéral et du cortex visuel primaire. Plusieurs modèles ont montré que cette propriété repose sur des principes auto-organisationnels induits par les stimuli (voir par exemple Linsker, 1986). Cependant, ces modèles demeurent souvent descriptifs et ne permettent pas d'aborder les raisons de fond qui président à l'organisation du système visuel. En d'autres termes, ils ne constituent pas une théorie explicative de cette organisation.

2.2. Modèles théoriques et optimisation du codage

Les considérations récentes qui ont conduit à mettre en évidence des éléments d'une théorie explicative de la vision sont essentiellement de nature écologique. Une première approche explicative des fonctionnalités visuelles s'appuie sur une série d'hypothèses sur la finalité évolutive des premières étapes de codage. A partir de principes informationnels, on peut s'intéresser à l'efficacité du codage mis en œuvre dans le système visuel. On est alors conduit à considérer ce principe d'efficacité comme un principe d'optimisation permettant de prédire l'agencement du système (Atick & Redlich, 1990).

Cette recherche d'optimalité porte également sur le caractère contraint, par l'organisation physique du monde, des images naturelles et leurs propriétés de symétrie et d'invariance (Turiel & Parga, 2003). On doit à Attneave (1954) et à Barlow (1983) la formulation de l'idée que l'une des finalités essentielles du codage visuel est de réduire la redondance des images. En présence d'une image naturelle, les réponses des récepteurs rétiniens, de même que les pixels d'une image numérique, sont en effet corrélés et présentent une forte redondance. La distribution statistique des pixels des images naturelles suit une loi d'échelle en $1/f^{\alpha}$ (Field, 1989), où f est la fréquence spatiale de l'image, caractéristique des structures autosimilaires. Un codage optimal des scènes capitalisant sur la nature de la distribution statistique des pixels est ainsi possible. Un tel codage maximisant l'indépendance statistique entre ses composantes a été recherché.

Olshausen et Field (1996) puis Bell et Sejnowski (1997) montrent que les filtres obtenus ont des profils similaires à des filtres d'orientation de Gabor, eux-mêmes assez similaires aux profils de réponse des champs récepteurs des cellules simples. Ainsi les contours à toutes les échelles sont les éléments essentiels supportant la structure des images naturelles. La statistique de ces images est alors dominée par ces événements rares. Les détecteurs obtenus sont locaux, fournissant une explication à l'existence de champs récepteurs et plus généralement à la rétinotopie de V1. Ils sont équivalents à des transformations en ondelette et les coefficients d'ondelettes obtenus constituent un codage clairsemé. Un tel codage a de nombreux avantages tels que la facilitation du codage dans les mémoires associatives, la

lisibilité accrue du code par les autres régions cérébrales et l'économie d'énergie. Olshausen et Field (1997) soulignent cependant le fait que la base d'ondelettes réellement utilisée par V1 est surcomplète, ce qui signifie que V1 réintroduit une certaine redondance permettant l'adaptation fine du codage à des situations variées.

3. La reconnaissance d'objets et de scènes

En vision computationnelle, de nombreuses méthodes ont été testées pour la reconnaissance de scènes et d'objets. La recherche s'oriente actuellement vers le codage des éléments à reconnaître dans une scène à l'aide de détecteurs invariants de bas niveau.

3.1. Détecteurs invariants et vision computationnelle

L'approche variationnelle du traitement d'images (Koenderink, 1984 ; Perona & Malik, 1990) met l'accent sur l'importance des gradients et d'une description multi-échelle dans l'analyse des images. Les implémentations qui en découlent diffèrent des implémentations biologiques mais recouvrent des processus de même ordre. Au-delà de la recherche des éléments les plus pertinents du codage que sont les contours, la description d'une scène nécessite en effet l'existence de représentations invariantes qui vont permettre de coder un objet à différentes échelles, à différentes positions dans le champ visuel et selon différentes vues.

Afin de pallier la sensibilité à l'occlusion, au changement d'échelle et à la variabilité intrinsèque des vues des modèles fondés sur l'apparence, des travaux ont cherché à mettre au point des détecteurs invariants en particulier en échelle et en rotation (Schiele & Crowley, 2000). L'indépendance entre descripteurs évite le recours à des relations spatiales sensibles à la pose et à l'apparence. Ainsi, un ensemble de descripteurs purement locaux permet d'obtenir une description des images suffisante pour établir une correspondance entre objets appris et parties de scènes visuelles (Lowe, 2004).

La plupart des approches considèrent les gradients comme des descripteurs essentiels (détecteurs SIFT par exemple ; Lowe, 2004), en accord avec les modèles précédents qui concluaient que les contours sont les composantes indépendantes des images naturelles (Bell & Sejnowski, 1997). Un aboutissement actuel de ces approches est la technique dite de « bag of words » (Lazebnik, Schmid, & Ponce, 2006) dans laquelle images et objets sont caractérisés par une collection de traits, par exemple ceux qui peuvent être capturés par des histogrammes multidimensionnels. Ces approches ne font cependant intervenir qu'un seul niveau de codage, ce qui n'est pas en accord avec l'organisation hiérarchique de la voie temporale. Elle ne tient également pas compte des relations spatiales entre les caractéristiques ni plus généralement des relations syntaxiques entre les éléments. On peut ainsi s'attendre à ce qu'elles ne capturent que très partiellement les caractéristiques complexes des objets et des scènes visuelles. Elles ont été mises en défaut, quelquefois par leurs auteurs eux-mêmes (Lazebnik, Schmid, & Ponce, 2006), en arguant du fait que des informations purement locales ne sont pas toujours suffisantes pour reconnaître un objet.

3.2. Modèles empiriques

De nombreux modèles empiriques ont tenté de rendre compte de l'aptitude de la perception à opérer de façon indépendante de l'orientation 3D de l'objet et également de la capacité à apprendre rapidement la forme d'un objet à partir d'un petit nombre de vues. Récemment Serre, Wolf, Bileschi, Riesenhuber et Poggio (2007) ont proposé un modèle capable d'apprendre à reconnaître aussi bien des objets et des catégories d'objets que des scènes

visuelles. Ce modèle met en avant l'existence plausible d'un code universel redondant. Récemment cette approche a été améliorée en prenant en compte le caractère clairsemé du codage (Mutch & Lowe, 2008).

Deux conceptions différentes de la hiérarchie des caractéristiques s'affrontent donc. Selon l'une, les caractéristiques intermédiaires sont constituées de détecteurs de parties d'objets portant une valeur sémantique (Dorkó & Schmid, 2003) alors que, selon l'autre, ils consistent en la combinaison non-linéaire de détecteurs du niveau précédent sans signification sémantique particulière (Ranzato, Huang, Boreau, & LeCun, 2007 ; Cadieu et al., 2007). La première approche vise à coder les relations qu'un objet contracte avec ses parties. Toutefois, elle renouvelle l'hypothèse du neurone grand-mère qui n'est pas acceptable lorsque le nombre d'objets croit de façon importante et ne rend pas compte de notre capacité d'interaction avec des objets nouveaux. Dans la seconde hypothèse l'existence d'un code redondant de haut niveau invariant en translation et en échelle permet l'apprentissage d'un nombre quelconque d'objets et de scènes. Il n'en reste pas moins que la façon dont ces détecteurs se combinent pour rendre compte des informations relationnelles entre les éléments d'un objet ou d'une scène reste largement ouverte.

3.3. Modèles théoriques

Une approche récente s'appuie sur des perceptrons multicouches possédant un nombre élevé de couches intermédiaires (Hinton, 2002). Ces réseaux « profonds » considèrent ainsi un système de codage formé d'un réseau de neurones formels organisé en une série de couches selon une architecture « feedforward » avec la propriété de reproduire à chaque couche les informations de la couche précédente.

Le principe inductif présidant à l'apprentissage dans ces réseaux conduit à minimiser la différence entre l'entrée native et l'entrée reconstruite. Hinton (2002) a développé un algorithme d'apprentissage rapide pour ce type de réseau ouvrant ainsi la voie à l'apprentissage de réseaux profonds multicouches dans des situations réalistes. Ce type d'approche peut être utilisé pour apprendre, à partir d'images, une hiérarchie de caractéristiques clairsemées utilisable pour la reconnaissance d'objets (Ranzato, Huang, Boreau, & LeCun, 2007). Une approche similaire (Lee, H., Ekanadham, & Ng, 2008) conduit à des détecteurs modélisant de façon assez fidèle les détecteurs d'orientation et de courbure de V1 et de V2. On retrouve dans ces modèles les ingrédients essentiels de la théorie du codage développée précédemment : indépendance et localité des descripteurs, caractère clairsemé du code mais également codage hiérarchique des caractéristiques de la scène. Le rapprochement entre ces modèles et les modèles empiriques décrits précédemment pourrait ouvrir la voie à une théorie générale du codage perceptif et de la reconnaissance d'objets.

Une autre classe de modèles cherche à capturer les fonctionnalités essentielles du système visuel par une approche bayesienne générique sans référence à son organisation. Ces modèles sont aptes à capturer l'essence des phénomènes sans toutefois avoir la capacité explicative des modèles précédents. Déjà selon Helmholtz (1867), la perception résulterait d'un équilibre entre expérience statistique a priori des structures présentes dans le monde et informations sensorielles. De tels modèles permettent la prise en compte d'informations contextuelles, depuis le contexte spatial - l'environnement visuel qui entoure immédiatement la vue d'un objet (Torralba, 2003) - jusqu'à des connaissances a priori concernant la nature de la tâche à réaliser, l'histoire personnelle de l'observateur ou les caractéristiques du monde dans lequel il évolue. La projection rétinienne bidimensionnelle d'objets par nature tridimensionnels génère par exemple des ambiguïtés qui ne peuvent être levées qu'à l'aide de connaissances ne

provenant pas de la scène elle-même. Le processus d'inférence bayésienne modélise bien ce genre de situations en considérant la perception visuelle comme un problème mal posé qui nécessite le recours à des connaissances préalables afin de contraindre l'unicité de la solution (Kersten, Mamassian, & Yuille, 2004). Ces modèles n'expliquent cependant pas l'implémentation des mécanismes qu'ils modélisent. Des travaux récents proposent des implémentations neuronales des processus bayesiens qui seraient à l'œuvre dans le système nerveux, visant ainsi à faire un lien entre modèles phénoménologiques et modèles explicatifs (voir Knill & Pouget, 2004).

Les modèles de reconnaissance d'objets précédemment décrits n'envisagent pas le recours à une phase attentionnelle. Comme nous l'avons mentionné, de ce point de vue, ils sont purement locaux et ne font intervenir aucune ségrégation entre les éléments de la forme et les éléments du fond. Un mécanisme supplémentaire est nécessaire pour réaliser cette ségrégation et le liage des éléments spécifiques afin de faciliter la reconnaissance d'objets. L'objectif de la section suivante est de présenter ces différents mécanismes attentionnels et leur rôle dans la perception visuelle.

4. L'attention

4.1. Modèles pré-attentionnels

Bien que le système visuel ait la capacité de traiter en parallèle l'information qui lui parvient, la recherche d'un objet particulier dans une scène résulte d'un processus décisionnel à l'issue d'une phase d'exploration. Ce processus est influencé par la tâche à réaliser et par le contexte (Yarbus, 1967). L'un des points essentiel est de savoir pourquoi et comment le système visuel réduit cette phase d'exploration de la scène. Deux points sont à prendre en considération : la capacité, éventuellement limitée, du système visuel à traiter l'information et la complexité intrinsèque de la tâche de recherche à réaliser pour identifier dans la scène les composantes en relation avec les attentes de l'observateur et les buts qu'il poursuit.

Au sein même du processus visuel, on distingue la pré-attention, processus ascendant guidé par les données et fondé sur la saillance intrinsèque des stimuli, et l'attention proprement dite, processus descendant modulé par le contexte, les attentes et les buts de l'observateur. Le terme d'attention sélective distingue l'ensemble des processus attentionnels des mécanismes liés de façon non spécifique à la notion d'éveil. On distingue par ailleurs un type d'attention dans laquelle l'élément saillant est une localisation (attention spatiale) d'un type d'attention dans laquelle l'élément saillant est associé aux caractéristiques de l'objet. D'un point de vue fonctionnel, on observe des modulations attentionnelles de l'activité des neurones dans l'aire V4 (Moran & Desimone, 1985), mais également dans V2 et V1 (Motter, 1993). La voie dorsale est également affectée par ces modulations en relation avec l'attention spatiale (Rushworth, Nixon, Renowden, Wade, & Passingham, 1997 ; voir aussi Itti, Rees, & Tsotsos, 2005, pour une revue récente des mécanismes attentionnels).

On doit à Treisman (1988) une théorie complète de l'attention sélective fondée sur l'idée que le système visuel code les stimuli selon des caractéristiques élémentaires similaires aux codages précoces observés dans le cortex visuel (couleur, orientation, disparité binoculaire et mouvement). Le modèle dérivé postule l'existence d'une carte de saillance intégrant ces caractéristiques. L'observation, pour des cibles définies par une conjonction de caractéristiques simples, d'une dépendance linéaire du temps d'identification par rapport au nombre de distracteurs a conduit à la métaphore du projecteur de l'attention, un processus qui examinerait de façon séquentielle les positions de la carte de saillance à la recherche de la

cible. Cette théorie est à l'heure actuelle la théorie dominante en ce qui concerne l'attention spatiale.

Parallèlement aux travaux des psychologues et des neurobiologistes, au début des années 90, plusieurs chercheurs en vision artificielle mettent en avant le concept de vision active (Aloimonos, 1993 ; Bajcsy, 1988) et construisent des dispositifs constitués de caméras mobiles pouvant rechercher activement une information dans une scène visuelle. Dès lors va se poser la question d'une définition formelle de la saillance, ces dispositifs étant essentiellement guidés par la saillance de la scène. Les définitions envisagées sont très nombreuses mais bon nombre d'entre elles demeurent vagues (voir Itti, Rees, & Tsotsos, 2005, pour une revue).

Le modèle pré-attentionnel actuellement le plus populaire est celui de Itti et Koch (2000). Il implémente l'essentiel des éléments de la théorie de l'intégration des caractéristiques. Cependant dans ce modèle, les cibles sont identifiées de façon passive, leurs caractéristiques ne prenant pas de part active au calcul de la saillance. La stratégie d'exploration d'une scène diffère ainsi très rapidement de la stratégie observée chez des humains qui explorent la scène selon ses caractéristiques sémantiques, soulignant la nécessité de prendre en compte les informations descendantes. Le modèle de Wolfe, Cave et Franzel (1989) est l'une des premières tentatives pour modéliser l'attention visuelle descendante. Il prend en compte les caractéristiques de la cible cherchée pour biaiser un modèle ascendant de calcul de saillance fondé sur la théorie de Treisman.

4.2. Le rôle de l'attention dans la reconnaissance d'objets

Dans une tentative pour lier pré-attention et sélection sémantique, Machrouh, Lienard et Tarroux (2001) montrent que les points de focalisation obtenus à l'aide d'un modèle ascendant permettent de sélectionner des régions d'une scène selon leurs caractéristiques sémantiques lorsque les instances de ces régions présentent des caractéristiques visuelles voisines. De leur côté, Walther et Koch (2006) cherchent à montrer comment un modèle ascendant permet de sélectionner des régions d'intérêt ciblées sur des objets potentiels (proto-objets). Afin de proposer un mécanisme d'implémentation plausible des processus attentionnels Niebur et Koch (1994) ont développé un modèle neuronal fondé sur la synchronisation des activités dans le cortex V4 qui reproduit quantitativement les expériences de Moran et Desimone (1985). La sélection de la localisation s'appuie sur une carte de saillance construite à partir des oppositions de couleur (V1, V2), de l'orientation (V1) et du mouvement (V3). Les auteurs proposent que cette carte se projette sur le cortex V4 permettant ainsi à l'attention spatiale qu'elle implémente de favoriser le traitement des formes présentes à la position correspondant à la saillance maximale.

Les modèles précédents se focalisent essentiellement sur l'attention spatiale et sa capacité à favoriser les objets présents à une localisation particulière. Une autre forme d'attention, en relation avec les caractéristiques des objets, semble moduler les neurones du cortex visuel (Maunsell & Treue, 2006). Dans ce mode, l'activité des neurones sensibles à des caractéristiques voisines de celles d'un stimulus situé à une position spatiale attendue est augmentée. Divers modèles tentent de combiner les caractéristiques des objets à rechercher avec la sélection de localisations préférentielles où chercher ces objets (voir Hamker, 2004). Pour une revue, on pourra consulter Koch (2004) et également Mozer et Vecera (2005) qui proposent un modèle unifié de l'attention spatiale et de l'attention aux objets et montrent que des résultats expérimentaux sont en faveur d'un modèle attentionnel de liage au bas niveau des caractéristiques élémentaires.

Partant de la constatation que le contexte peut être une source riche d'informations perceptives Torralba (2003) a proposé un modèle probabiliste qui montre comment la connaissance du contexte visuel immédiat peut être utilisée pour faciliter la recherche et la reconnaissance d'un objet. De telles connaissances contextuelles sont alors mises à profit pour diriger l'attention vers les régions de la scène où la présence de l'objet est la plus probable. Le modèle est en accord avec les fixations observées chez des observateurs humains recherchant le même type d'objet.

4.3. Rôle et justification de l'existence de processus attentionnels

La première justification de l'existence de mécanismes attentionnels est la supposée limitation de capacité du système visuel qui imposerait un choix parmi les stimuli pour éviter une saturation (Broadbent, 1971). Parmi les premiers à avoir remis en cause cette vision, Allport (1989) propose que les mécanismes attentionnels ont évolué pour satisfaire un ensemble de finalités biologiques ou computationnelles plutôt que sous la contrainte d'une trop grande quantité d'information à traiter. Allport oppose ainsi à la théorie de Broadbent une théorie qui place l'attention et le besoin d'un système de sélection au niveau du contrôleur (entre perception et action par exemple). Cette vision des choses, si elle résidait seulement sur la sélection, à un niveau post-sémantique, d'une partie de l'information pourrait encore s'expliquer par une contrainte informationnelle.

Allport conçoit en fait ce mécanisme comme le résultat d'une sélection positive par l'évolution et non comme le résultat d'une limitation. Pour Allport, la cohérence comportementale est la finalité principale du processus attentionnel. Un comportement cohérent implique l'assignation de priorités et une coordination à différents niveaux (motivationnels, cognitifs, moteur, sensoriel). Dans ces conditions, le processus attentionnel, élément essentiel de la cohérence comportementale, peut être la cible de la sélection naturelle. Il ne s'agit plus d'un processus imposé par une quelconque limitation interne du système sans référence à l'environnement sensoriel et comportemental, mais d'une aptitude incarnée en relation étroite avec la spécification des buts et des motivations de l'individu.

En accord avec cette théorie, Rimey et Brown (1992) proposent un modèle de tâche orientée vision construit à l'aide de réseaux bayesiens. Il permet de sélectionner, d'une façon orientée par la tâche, les éléments pertinents pour l'action en cours. Ce modèle est à rapprocher des observations de Hayhoe, Droll et Mennie (2007) qui montrent une modulation des fixations attentionnelles en fonction de la tâche en cours dans une tâche similaire à celle utilisée par Rimey et Brown (localisation d'éléments du couvert sur une table). Selon une autre approche (Tsotsos, 1992), c'est la complexité intrinsèque des tâches visuelles qui conduit à justifier l'existence de processus attentionnels comme étant le seul moyen de résoudre un problème perceptif par ailleurs intraitable au sens computationnel. Tsotsos montre que la résolution du problème passe par la limitation de l'exploration de la scène à un petit nombre de positions, justifiant ainsi les mécanismes attentionnels sans pour autant faire appel à une hypothèse de limitation de capacité. Il en est de même si ce sont les caractéristiques de l'objet qui servent à le rechercher.

Dans la ligne de cette analyse en termes de complexité, Tsotsos et al. (1995) ont proposé un modèle fondé sur la notion de « selective tuning ». Le modèle reprend l'analyse théorique précédente pour conclure à la nécessité d'une sélection des caractéristiques à rechercher et à la définition de régions d'intérêt. Il introduit également l'idée d'une hiérarchie de traitements nécessaires pour simplifier le traitement du flux informationnel ainsi que l'idée qu'un processus attentionnel arbitre en permanence entre un processus dirigé par les données et un

processus orienté par la tâche. Cet arbitrage est réalisé par la mise en œuvre de deux flux d'information, une pyramide de traitements ascendants et une hiérarchie de winner-takes-all descendante. La recherche d'un objet guidée par l'attention implémenterait ainsi un processus inférentiel d'hypothèse et de test qui, comme dans la proposition d'Allport, peut être la cible de l'évolution. Considérer que l'attention est au centre d'un mécanisme d'inférence dont la finalité essentielle est la constitution puis la validation d'hypothèses d'interprétation de la scène observée rejoint encore une fois l'intuition de Helmholtz (1867) proposant de considérer la perception comme un processus inférentiel.

La plupart des modèles précédents sont mis en œuvre dans des tâches de reconnaissance statique où la relation entre perception et action ne joue pas de rôle. Des approches récentes montrent que le rôle de l'attention devient crucial dans des tâches qui supposent une organisation temporelle de l'acquisition des informations visuelles. C'est en particulier le cas des modèles qui visent une implémentation robotique (Lee, Buxton, & Feng, 2005) ou plus généralement qui visent à rendre compte de l'usage de l'attention dans un contexte comportemental et écologique en relation avec une tâche particulière (Hayhoe, Droll, & Mennie, 2007). Ces observations et ces modèles privilégient ainsi la relation entre l'ordre temporel des fixations oculaires et la planification motrice. En permettant un couplage harmonieux entre information sensorielle et planification motrice, le rôle de l'attention dépasse bien la question de la limitation de capacité. Le processus attentionnel s'inscrit ainsi dans le processus dynamique induit par le comportement et participe à la cohérence de ce comportement.

5. Perception et action

5.1. Perception pour l'action

Dans une conception située, perception et action sont intimement liés. La perception fournit des éléments pour l'action et en retour, la perception est influencée par l'action. Dans ce contexte, on va dans un premier temps examiner les modèles qui permettent de sélectionner l'action à réaliser en fonction des perceptions que le système reçoit puis ceux qui permettent de construire une représentation de l'environnement au cours de l'action. Compte tenu des incertitudes de perception et de position, les modèles appropriés sont encore une fois des modèles probabilistes. Cette classe de modèles s'étend des simples modèles de décision à ceux qui construisent un système de sélection de l'action sur la base d'un apprentissage par renforcement dans des conditions d'observabilité limitée (Kaelbling, 1998).

O'Reilly (1996) propose de modéliser l'interaction perceptive avec le monde extérieur comme l'apprentissage d'un modèle interne du monde et en particulier d'un modèle d'anticipation à partir d'observations imparfaites sur l'état du monde. La perception est alors conçue comme un processus d'estimation et de prédiction de l'état de l'environnement à partir d'observations incomplètes. Les modèles qui visent, toujours sur la base d'observations incomplètes, à localiser un robot tout en construisant la carte de son environnement au cours d'une exploration (SLAM, Durrant-Whyte & Bailey, 2006) sont les héritiers directs de ces propositions. Au-delà de la simple localisation spatiale à partir d'amers visuels qui conduit à établir une cartographie de l'environnement se pose la question de l'identification des lieux visités. Là encore les méthodes bayésiennes, éventuellement combinées à des mécanismes attentionnels, permettent de développer des modèles pertinents (Guillaume, 2009). Au-delà de ces réalisations, on peut désormais envisager la programmation complète d'un contrôleur robotique, depuis les mécanismes de filtrage et de codage de l'information jusqu'aux

processus de sélection de l'action, le tout incluant des possibilités d'apprentissage sous l'unique paradigme de l'approche bayesienne (Thrun, 2000).

5.2. Perception dans l'action

Avec l'objectif de défendre l'idée d'une perception active, Gibson (1966) détermine deux manières possibles de concevoir les sens. Soit les sens sont des canaux de sensations, essentiellement passifs, qui sont à l'origine des qualités de l'expérience. Soit ils sont des systèmes perceptifs, essentiellement actifs, qui extraient des informations sources de connaissances sur le monde. Pour Gibson, la perception visuelle n'est pas basée sur la sensation visuelle, mais sur l'information contenue dans les propriétés structurelles invariantes des stimuli : percevoir, c'est extraire, grâce aux mouvements, cette information en en détectant les invariants. L'hypothèse que les objets de notre perception ne sont pas à proprement parler les invariants de la sensation, mais plutôt les invariants de cercles sensorimoteurs inséparables de l'activité de l'observateur a été développée en psychologie (O'Regan & Noë, 2001 ; Varela, Thompson, & Rosch, 1991), en philosophie (Merleau-Ponty, 1945) et s'est étendue à de nombreux champs de recherches théoriques, scientifiques et technologiques tels que la cognition située, la robotique autonome (Brooks, 1991 ; Gaussier, Moga, Banquet, & Quoy, 1998) et l'ergonomie. Ainsi par exemple, la reconnaissance d'un objet serait fondée sur le rôle de cet objet comme attracteur d'une dynamique comportementale et non sur ses propriétés intrinsèques (Schöner, Dose, & Engels, 1995). Des modèles théoriques ont démontré que l'interaction avec le monde permet d'acquérir des connaissances sur celui-ci (Philipona, O'Regan, & Nadal, 2003). Ces positions rejoignent celles de Prochiantz (1997) selon qui le cerveau antérieur s'est développé en réponse au besoin de complexification de boucles sensori-motrices et celles de Berthoz (1997) qui met en avant la construction par l'action et dans l'action des capacités cognitives.

Cette conception de la perception comme constituée de boucles sensorimotrices fait écho à celle de Von Uexküll (1934) en éthologie. Pour cet auteur, il est possible de caractériser les « mondes propres » (*Umwelt*) de chaque être sur la base des boucles sensorimotrices qui constituent sa relation au monde. Il donne l'exemple du monde propre de la tique en se basant sur trois boucles sensorimotrices et souligne que ce monde propre se réduit à ces trois cercles fonctionnels. La caractérisation des mondes propres des individus sur la base des couplages sensori-moteurs qui caractérisent leur relation au monde a également été employée pour évoquer le monde propre des nourrissons (Stern, 1989).

5.3. Illustrations en psychologie expérimentale

Le lien indissociable entre la perception et l'action s'illustre notamment par la plasticité du système nerveux, sa capacité à modifier durablement sa propre structure en acquérant des possibilités nouvelles de fonctionnement. Les changements ont lieu sous l'effet de conditions imposées par l'environnement et ils peuvent s'observer soit lors de la maturation du système au cours de son développement, soit ultérieurement, lorsque la structure est stabilisée en fin de croissance.

Le rôle de la motricité de l'organisme sur la mise en place des structures de fonctionnement perceptif lors du développement s'illustre par les travaux de Held et Hein (1963). Les auteurs ont élevé des chatons dans l'obscurité pendant plusieurs semaines après la naissance. A partir de la quatrième semaine, les chatons sont placés par paire, trois heures par jour, dans un manège éclairé constitué visuellement de lignes verticales. Un des chatons est attelé à l'une des branches du manège et il entraîne par son mouvement locomoteur l'autre chaton suspendu

dans une nacelle à l'autre branche du manège. Les deux chatons ont donc la même expérience visuelle, sauf qu'elle est associée à une exploration active pour l'un et à un transport passif pour l'autre. Après une période d'habituation, le chaton actif présente un comportement visuo-moteur normal. En revanche, le chaton passif présente des déficits visuo-moteurs : il butte contre les obstacles, tombe à l'extrémité des tables et est incapable de diriger correctement le mouvement de ses pattes pour les poser sur un support solide. Ce chaton n'est pas à proprement parler aveugle : il est capable d'identifier des formes mais il est incapable de localiser ces formes dans un espace constitué. Le chaton passif n'a pas constitué son espace des lieux : il ne dispose pas des repères spatiaux nécessaires à l'orientation correcte de ses activités motrices. Le rôle de l'activité sur l'organisation de l'espace sensori-moteur s'est illustré par de nombreux travaux ultérieurs. Pour en donner un exemple, Hein et Held (1967) ont montré qu'un chaton empêché par une collerette opaque de voir ses membres antérieurs, construit correctement son espace locomoteur. Cependant, il se révèle incapable de guider visuellement le mouvement de ses pattes. En particulier, il ne présente pas d'extension réflexe des pattes avant lorsqu'on l'approche d'une surface d'appui.

Le rôle de la vision du corps en mouvement dans l'établissement de relations visuo-motrices coordonnées a aussi été étudié chez l'homme, notamment lors d'expériences d'adaptation prismatique. Lorsque l'on demande à un observateur adulte de pointer vers une cible visuelle alors que l'espace visuel est déplacé par le port de lunettes prismatiques, la personne manque la cible, puisque le prisme dévie les rayons lumineux. On permet ensuite pendant quelques minutes à la personne de déplacer activement sa main devant les prismes en contrôlant la position finale de sa main. Puis, lors de la phase de test, on lui demande de pointer de nouveau vers la cible sans le contrôle de la position finale de sa main, c'est-à-dire sans pouvoir apprécier son erreur éventuelle. La personne pointe alors correctement vers la cible ; elle a donc réajusté son programme moteur. En revanche si, au cours de la phase d'adaptation, au lieu d'un déplacement actif c'est l'expérimentateur qui déplace passivement la main de l'observateur, il n'y a pas de réorganisation des programmes moteurs. Lors de la phase de test l'observateur continue de pointer incorrectement vers la cible visuelle lorsqu'il ne peut voir la position finale de sa main (Paillard & Brouchon, 1968).

Une autre expérience intéressante d'adaptation a été mise au point par Kohler (1951) qui a porté un appareil optique qui inverse l'image rétinienne, de telle sorte que le monde apparaît inversé haut/bas et droite/gauche. Kohler et les personnes qui ont répliqué cette expérience se sont adaptés au bout de quelques jours. Après deux semaines, ils sentent que leur monde visuel est normal de nouveau. Ce qui est intéressant c'est que durant le cours de l'adaptation, la perception du monde est sujette à une sorte de fragmentation et à une dépendance au contexte et à la tache. Certaines choses sont redressées mais il demeure des ambiguïtés et des inconsistances, par exemple certains éléments tels que les plaques de voiture restent inversées. Cette expérience suggère que l'orientation et la localisation d'objets dans le champ visuel peut être définie par rapport à de multiples référents et de multiples tâches, et chaque tâche s'adapte indépendamment, en fonction des actions sur les objets perçus.

6. Perception visuelle et multisensorialité

6.1. Exemples en psychologie expérimentale

Jusqu'à présent, nous avons abordé les mécanismes de bas niveau et de haut niveau caractérisant la perception visuelle. Cependant, la vision ne doit pas être considérée comme étant isolée des autres modalités sensorielles. S'agissant de la mémoire à court terme, il a été

suggéré que les informations provenant d'une scène visuelle ne sont pas stockées pour une utilisation ultérieure (rappel ou comparaison) selon un format spécifique à la modalité de présentation des stimuli ; certaines propriétés sont extraites et représentées selon un format abstrait ou amodal (Irwin & Andrew, 1996). S'agissant des mécanismes attentionnels, de nombreuses études ont montré que l'attention spatiale est déterminée de manière multisensorielle. Par exemple, la présentation d'un stimulus dans une modalité sensorielle (e.g., le toucher) capture l'attention de manière exogène de telle sorte que cela facilite le traitement de stimuli provenant d'autres modalités sensorielles (e.g., la vision ou l'audition) présentés au même emplacement (voir Spence, McDonald, & Driver, 2004, pour une revue).

Les tâches de détection de changements, et en particulier le paradigme de la cécité aux changements intermodale, permettent de comprendre quelles sont les similitudes et les différences des mécanismes d'encodage de l'information présentée à travers différentes modalités sensorielles ainsi que le caractère multisensoriel de notre attention spatiale. Le phénomène de cécité aux changements se produit lorsqu'une perturbation introduite au sein de la scène perçue au moment du changement empêche les observateurs de détecter ce changement et ce, bien que cette modification puisse être d'importance et parfaitement détectable dans des conditions normales de perception. Ce phénomène a été mis en évidence pour les modalités visuelles, tactiles et auditives (e.g., Auvray & O'Regan, 2003 ; Gallace, Tan, & Spence, 2006 ; Vitevitch, 2003). Le phénomène de cécité aux changements se produit aussi lorsque les perturbations sont présentées dans une autre modalité sensorielle que le changement. En effet, les participants échouent à détecter la présence d'un changement de position entre deux scènes vibrotactiles présentées sur la surface du corps, non seulement lorsque des perturbations vibrotactiles sont utilisées pour masquer le changement, mais aussi lorsque des perturbations visuelles sont utilisées (Gallace, Auvray, Tan, & Spence, 2006). En revanche, dans les mêmes conditions de présentation, des distracteurs vibrotactiles n'entraînent pas de cécité aux changements visuels (Auvray, Gallace, Tan, & Spence, 2007) et des distracteurs auditifs n'entraînent pas de cécité aux changements tactiles (Auvray, Gallace, Hartcher-O'Brien, Tan, & Spence, 2008).

Des travaux ont aussi exploré la possibilité d'une cécité aux changements intermodale ; c'est-à-dire lorsque les deux ensembles de stimulation à comparer sont présentés dans deux modalités sensorielles différentes : l'une visuelle et l'autre tactile. Les résultats montrent qu'en l'absence de masque, les participants peuvent détecter avec précision les changements de position, en dépit du fait que les deux ensembles de stimulation à comparer soient présentés dans deux modalités sensorielles différentes. En revanche, lorsqu'un masque est introduit entre ces deux ensembles de stimulation, une cécité aux changements survient de manière similaire, que le masque soit visuel ou tactile (Auvray, Gallace, Tan, & Spence, 2007). En résumé, la possibilité de comparer des positions à travers différentes modalités sensorielles suggère que certaines des informations requises pour comparer des emplacements spatiaux sont stockées selon un format amodal. En revanche, les asymétries dégagées dans les mécanismes de détection de changements suggèrent que certaines des informations sont encodées selon un cadre de référence spécifique à chaque modalité sensorielle (par exemple, rétinotopique pour la vision).

6.2. Modèles

Les modèles formels de perception multimodale sont à mettre en relation avec la question centrale de la combinaison de capteurs en robotique. De nombreuses propositions ont été faites pour agréger les informations provenant de diverses sources de données. La prise en compte de l'incertitude liée au caractère bruité des capteurs a conduit aux approches

bayesiennes de fusion multi-modale (Singhal & Brown, 1997). Les principes de maximisation de l'information mutuelle permettent par exemple de lier les images d'un couple stéréoscopique. On peut également de cette façon fusionner des données visuelles et proprioceptives dans une tâche de manipulation d'objets guidée par la vision. Prodanov, Drygajlo, Richiardi et Alexander (2008) proposent un système d'interaction multimodale avec fusion d'informations provenant à la fois de détecteurs de présence, de la vision et d'une analyse de la parole pour un robot en interaction avec des humains réalisant des tâches de dialogue.

Avillac, Denève, Olivier, Pouget et Duhamel (2005) soulignent le fait que notre perception du monde est essentiellement multimodale, combinant par exemple des informations visuelles, auditives et proprioceptives. Ces modalités sensorielles se basent sur différents cadres de références. L'approche bayesienne ne prend pas en compte la diversité des cadres de références dans lesquels ces modalités s'expriment. En créant des liens cross-modaux dans les fonctions de base qu'ils utilisent pour rendre compte des cadres de références, ces auteurs montrent comment combiner multisensorialité et cadres de référence multiples. Leur modèle suggère que l'intégration multisensorielle est fondée sur un dialogue entre modalités plutôt que sur la convergence de l'ensemble des informations sensorielles dans un modèle unique.

7. Conclusion

7.1. Les trois approches de la modélisation

On voit ainsi se dessiner trois approches de la modélisation : théorique visant à révéler les principes fondamentaux sous-jacents au processus perceptif, empirique visant à comprendre l'implémentation de ces principes et phénoménologique visant à rendre compte de la complexité des processus. Les premiers modèles développés convergent vers l'existence d'un alphabet de formes mis en place durant le développement dont les combinaisons permettent le codage des formes qui nous entourent. Ils apportent des preuves de l'influence du contexte sous ses divers aspects dans le processus visuel. Les modèles théoriques ont ensuite permis de faire le lien avec les domaines de la modélisation et du traitement des données parfois a priori très éloignés (la séparation de source pour le codage ou la mécanique des fluides pour la théorie de l'espace échelle).

On assiste à un usage croissant de modèles bayesiens permettant de gérer les situations d'incertitude et l'incomplétude, de permettre une modélisation phénoménologique des processus malgré tout fondée sur des principes éprouvés et de produire des modèles robustes et simples d'utilisation. On est ainsi passé progressivement de modèles ponctuels de tel ou tel processus à une théorie générale de l'inférence, de l'apprentissage et de la programmation bayesienne. Ces modèles apportent un cadre de description et des possibilités de combinaison de l'information unifiés. Dans un contexte robotique, ils font échos aux approches bayesiennes de la localisation, de la sélection de l'action, de la planification et de la programmation robotique.

Une question reste cependant ouverte : le cerveau est-il bayesien ? En d'autres termes, ces modèles sont-ils une bonne représentation des processus mis en œuvre dans le système nerveux et, si tel est le cas, comment sont-ils implémentés ? Comme nous l'avons vu, des travaux récents visent à répondre à cette question (voir par exemple Doya, Ishii, Pouget, & Rao, 2007). Cependant, quel que soit le paradigme sous lequel on se place, le caractère situé de la perception ainsi que la construction de capacités cognitives dans une interaction constante avec le monde impose que le système qui met en œuvre le modèle soit en

interaction avec le monde. Par conséquent, seuls les modèles implémentés dans des systèmes robotiques permettent de mettre en œuvre de véritables capacités perceptives, de vérifier la pertinence des modèles et leur adaptation dans un environnement donné. Toute autre voie aboutit à fournir a priori les structures du monde et à les vider ainsi de leur sens.

7.2. La perception comme processus global et situé

On ne peut pas réduire la reconnaissance d'objets ou de scènes à un simple problème de catégorisation. La catégorie conceptuelle ne peut être donnée a priori car une véritable reconnaissance passe par le fait de reconnaître cette catégorie dans ses fonctionnalités et ses usages et pas seulement au travers des attributs visuels qui la caractérisent. Le cadre de l'apprentissage supervisé doit ainsi être dépassé pour placer le système en situation où le concept émerge selon des critères d'utilité et de fonctionnalité en relation avec une rétribution.

La perception est un état global qui ne peut être appréhendé que de façon holistique. Les éléments sur lesquels elle est fondée sont susceptibles d'être formalisés séparément mais la perception est le résultat de leur mise en commun adéquate au sein d'un système comportemental et non par l'intervention du raisonnement. Il y a donc bien une nécessaire unité de la perception qui ne se conçoit à la fois que dans son rapport à l'expérience sensible et en lien avec les différentes modalités sensorielles qui se rattachent à un perçu donné (Dokic, 2004). Merleau-Ponty (1945) souligne le caractère nécessairement situé de la perception. Selon lui, elle résulte d'un rapport au monde pas uniquement construit par l'entendement, mais construit par « un être qui y est jeté et qui y est attaché comme par un lien naturel ».

La conception interactionniste et constructiviste de la perception conduit à considérer que les fonctionnalités des systèmes perceptifs ont émergé du processus évolutif à l'issue de mécanismes d'optimisation. C'est en particulier visible lorsqu'on analyse la façon dont le cortex visuel primaire code l'information qui lui parvient. Toutefois ce processus d'optimisation a la particularité d'être sous-tendu par un processus évolutif qui ne génère pas des solutions à des problèmes posés a priori mais adapte des solutions découvertes par hasard à des situations particulières. La multiplicité des solutions ainsi produites démontre qu'il ne faut pas chercher dans la perception ou dans les capacités cognitives la solution à un problème général. De ce point de vue la tique de von Uexküll est une solution tout aussi satisfaisante que nos cerveaux capables de performances cognitives élaborées. L' « Umwelt » est spécifique de l'espèce. Ainsi, c'est en s'interrogeant sur la nature des problèmes qu'un système donné est capable de résoudre plutôt que sur la façon de construire un système en vue de résoudre un problème donné a priori qu'on parviendra à comprendre quelles dispositions sont nécessaires en face de telle ou telle contrainte écologique.

8. Bibliographie

Adelson, E. H., & Bergen, J. R. (1985). Spatio-temporal energy models for the perception of motion. *Journal of the Optical Society of America, A2,* 284-299.

Agre, P. E. (1988). *The dynamic structure of everyday life.* Unpublished PhD Thesis, MIT, Boston.

Allport, A. (1989). Visual attention. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 631-682). Boston, MA: MIT Press.

Aloimonos, Y. (Ed.). (1993). *Active perception*. Hillsdale, NJ: Lawrence Erlbaum.

Atick, J. J., & Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computation, 2*, 308-320.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*, 183-193.

Auvray, M., Gallace, A., Hartcher-O'Brien, J., Tan, H. Z., & Spence, C. (2008). Tactile and visual distractors induce change blindness for tactile stimuli presented on the fingertips. *Brain Research, 1213*, 111-119.

Auvray, M., Gallace, A., Tan, H. Z., & Spence, C. (2007). Crossmodal change blindness between vision and touch. *Acta Psychologica, 126*, 79-97.

Auvray, M., & O'Regan, J. K. (2003). L'influence des facteurs sémantiques sur la cécité aux changements progressifs dans les scènes visuelles. *Année Psychologique, 103*, 9-32.

Avillac, M., Denève, S., Olivier, E., Pouget, A., & Duhamel, J. R. (2005). Reference frames for representing visual and tactile locations in parietal cortex. *Nature Neuroscience, 8*, 941-949.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE, 76,* 996-1005.

Barlow, H. B. (1983). Understanding natural vision. In O. J. Braddick & A. C. Sleigh (Eds.), *Physical and Biological Processing of Images* (Vol. 11, pp. 2-14). Berlin: Springer-Verlag.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research, 37*, 3327-3338.

Berthoz, A. (1997). *Le sens du mouvement*. Paris: Odile Jacob.

Broadbent, D. E. (1971). *Decision and stress*. London: Pergamon.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*, 139-159.

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C. E., Riesenhuber, M., & Poggio, T. (2007). A model of V4 shape selectivity and invariance. *Journal of Neurophysiology, 98*, 1733-1750.

Dokic, J. (2004). *Qu'est-ce que la perception?* Paris: Vrin.

Dorkó, G., & Schmid, C. (2003). Selection of scale-invariant parts for object class recognition. *IEEE International Conference on Computer Vision (ICCV)*, 634-640.

Doya, K., Ishii, S., Pouget, A., & Rao, A. (Eds.). (2007). *Bayesian brain. Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localisation and mapping (SLAM): Part I the essential algorithms. *Robotics and Automation Magazine, 13*, 99-110.

Ewert, J. P., & Arbib, M. A. (1989). *Visuomotor coordination: Amphibian, comparisons, models, and robots*. New York: Plenum Press.

Field, D. J. (1989). What the statistics of natural images tell us about visual coding. *SPIE, 1077*, 269-276.

Gallace, A., Auvray, M., Tan, H. Z., & Spence, C. (2006). When visual transients impair tactile change detection: A novel case of crossmodal change blindness? *Neuroscience Letters, 398*, 280-285.

Gallace, A., Tan, H. Z., & Spence, C. (2006). The failure to detect tactile change: A tactile analog of visual change blindness. *Psychonomic Bulletin & Review, 13*, 300-303.

Gaussier, P., Moga, S., Banquet, J. P., & Quoy, M. (1998). From perception-action loops to imitation processes: A bottom-up approach of learning by imitation. *Applied Artificial Intelligence: An International Journal, 7*, 701-727.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Guillaume, H. (2009). *Explorer pour reconnaître : Approche probabiliste pour la reconnaissance visuelle de lieux par un robot mobile.* Unpublished PhD Thesis, Paris Sud 11.

Hamker, F. H. (2004). A dynamic model of how feature cues guide spatial attention. *Vision Research, 44*, 501-521.

Harnad, S. (1990). The symbol grounding problem. *Physica, 42, 335-346.*

Hayhoe, M., Droll, J., & Mennie, N. (2007). Learning where to look. In R. P. G. Van Gompel, M. H. Fischer, W. S. Murray & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 641-660). Oxford: Elsevier.

Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (1996). Computational models of cortical visual processing. *Proceedings of the National Academy of Sciences of the United States of America, 93*, 623-627.

Hein, A., & Held, R. (1967). Dissociation of the visual placing response into elicited and guided components. *Science, 158*, 390-392.

Held, R., & Hein, A. (1963). Movement produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology, 56*, 872-876.

Helmholtz, H.v. (1867). *Handbuch der physiologischen Optik*, Leipzig: Leopold Voss.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation, 14*, 1771-1800.

Hübel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interactions, and functional architecture in the cat's visual cortex. *Journal of Physiology, 160*, 106-154.

Irwin, D. E., & Andrew, R. V. (1996). Integration and accumulation of information across saccadic eye movements. In T. Inui & J. L. McClelland (Eds.), *Attention and performance* (Vol. XVI – Information integration in perception and communication, pp. 125-155). Cambridge, MA: MIT Press.

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

Itti, L., Rees, G., & Tsotsos, J. (Eds.). (2005). *Neurobiology of Attention*, Los Angeles, CA: Elsevier.

Kaelbling, L. P. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence, 101*, 99-134.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271-304.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*, 712-719.

Koch, C. (2004). Selective visual attention and computational models. *CNSBi*, *186*, 1-14.

Koenderink, J. J. (1984). The structure of images. *Biological Cybernetics, 50*, 363-370.

Kohler, I. (1951) Über aufbau und wandlungen der wahmehmungswelt. *Österreichische Akademie der Wissenschaften. Sitzungsberichte,philosophisch-historische Klasse, 227*, 1-118.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2*, 2169-2178.

Lee, H., Ekanadham, C., & Ng, A. Y. (2008). Sparse deep belief net model for visual area V2. *Neural Information Processing Systems (NIPS), 20,* 873-880.

Lee, K., Buxton, H., & Feng, J. (2005). Cue-guided search: A computational model of selective attention. *IEEE Transactions on Neural Networks, 16*, 910-924.

Linsker, R. (1986). From basic network principles to neural architecture: Emergence of orientation-selective cells. *Proceedings of the National Academy of Sciences of the United States, 83*, 8390-8394.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*, 91-110.

Machrouh, Y., Lienard, J. S., & Tarroux, P. (2001). Multiscale feature extraction from the visual environment in an active vision system. *International Workshop on Visual Form 4*, 1-5.

Maunsell, J. H. R., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences, 29*, 317-322.

Merleau-Ponty, M. (1945). *La Phénoménologie de la Perception*. Paris: Gallimard.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science, 229*, 782-784.

Motter, B. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2 and V4 in the presence of competing stimuli. *Journal of Neurophysiology, 70*, 909-919.

Mozer, M. C., & Vecera, S. P. (2005). Object- and space-based attention. In L. Itti, G. Rees & J. Tsotsos (Eds.), *Neurobiology of attention* (pp. 130-134). New York: Elsevier.

Mutch, J., & Lowe, D. G. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision, 80*, 45-57.

Niebur, E., & Koch, C. (1994). A model for the neuronal implementation of selective visual attention based on temporal correlation among neurons. *Journal of Computational Neuroscience, 1*, 141-158.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*, 883-917.

O'Reilly, R. C. (1996). *The leabra model of neural interactions and learning in the neocortex.* Unpublished PhD Thesis, Carnegie Mellon University, Pittsburgh, PA.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607-609.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*, 3311-3325.

Paillard J., & Brouchon, M. (1968). Active and passive movements in the calibration of position sense. In S. J. Freedman, *The Neuropsychology of spatially oriented behavior* (pp. 37-55). Homewood III., Dorsey Press.

Perona, P., & Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*, 629-639.

Philipona, D., O'Regan, J. K., & Nadal, J. P. (2003). Is there something out there? Inferring space from sensorimotor dependencies. *Neural Computation, 15*, 2029-2050.

Prochiantz, A. (1997). *Les anatomies de la pensée*. Paris: Odile Jacob.

Prodanov, P., Drygajlo, A., Richiardi, J., & Alexander, A. (2008). Low-level grounding in a multimodal mobile service robot conversational system using graphical models. *International Service Robotics, 1*, 3-26.

Ranzato, M., Huang, F.-J., Boreau, Y., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-8.

Rimey, R., & Brown, C. (1992). Task-oriented vision with multiple Bayes nets. In A. Blake & A. Yuille (Eds.), *Active vision* (pp. 217-238). Cambridge, MA: The MIT Press.

Rushworth, M. F. S., Nixon, P. D., Renowden, S., Wade, D. T., & Passingham, R. E. (1997). The left parietal cortex and motor attention. *Neuropsychologia, 35*, 1261-1273.

Schiele, B., & Crowley, J. L. (2000). Recognition without correspondance using multidimensional receptive field histograms. *International Journal of Computer Vision, 36*, 31-50.

Schöner, G., Dose, M., & Engels, C. (1995). Dynamics of behavior: Theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems, 16*, 213-245.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*, 411-426.

Singhal, A., & Brown, C. (1997). Dynamic Bayes net approach to multimodal sensor fusion. *SPIE, 3209*, 2-10.

Spence, C., McDonald, J., & Driver, J. (2004). Exogenous spatial-cueing studies of human cross-modal attention and multisensory integration. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 277-320). Oxford, UK: Oxford University Press.

Stern, D. (1989). *Le monde interpersonnel du nourisson*. Paris: PUF.

Thrun, S. (2000). Towards programming tools for robots that integrate probabilistic computation and learning. *IEEE International Conference on Robotics and Automation (ICRA),* 306-312.

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer Vision, 53*, 169-191.

Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology, 40A*, 201-237.

Tsotsos, J. K. (1992). On the relative complexity of active v.s. passive visual search. *International Journal of computer vision, 7*, 127-141.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*, 507-545.

Turiel, A., & Parga, N. (2003). Role of statistical symmetries in sensory coding: An optimal scale invariant code for vision. *Journal of Physiology, 97*, 491-502.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale & R. J. W. Mansfield (Eds.). *Analysis of visual behavior* (pp. 549-586). Cambridge, MA: MIT Press.

Van Essen, D. C., Anderson, C. H., & Felleman, D. J. (1992). Information processing in the primate visual system: An integrated systems perspective. *Science, 255*, 419-423.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 333-342.

Von Uexküll, J. (1934). *Mondes animaux et monde humain (1965 ed.).* Paris: Denoël.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395-1407.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology, 15*, 419-433.

Yarbus, A. L. (1967). *Eye movements and vision* (B. Haigh, Trans.). New York: Plenum Press.