De la Reconnaissance des Objets à la Reconnaissance des Visages

Hervé Abdi & Dominique Valentin

The University of Texas at Dallas & Université de Bourgogne à Dijon

Résumé

Nous exposons quelques modèles récents de la reconnaissance des objets et des visages. Dans un premier temps, nous décrivons la théorie des *géons* de Biederman qui s'applique à la reconnaissance des objets. Cette théorie suppose, pour l'essentiel, que les objets sont analysés par le système visuel en un ensemble de formes élémentaires (les géons) dont la composition spécifie chaque objet. Cette théorie permet de rendre compte d'un nombre important de données empiriques, mais, récemment certains de ses présupposés se sont trouvés remis en cause. Les visages constituent clairement une exception pour le modèle de Biederman (puisque les visages possèdent tous la même composition géonique). Les modèles connexionistes semblent particulièrement adaptés pour ces stimuli. Un problème, toutefois, pour ces modèles reste le développement d'une représentation tri-dimensionelle. Nous décrivons, en conclusion, quelques résultats récents suggérant que cette représentation peut apparaître spontanément comme une conséquence du stockage en mémoire d'images de visages.

1. Introduction

Dans ce chapitre nous évoquons quelques modèles récents consacrés à la reconnaissance des objets et des visages. Il peut sembler curieux de séparer ces deux domaines. Après tout, on pourrait considérer que les visages ne sont que des objets parmi d'autres. Ce parti pris se défend mal si l'on considère, en particulier, les données récentes de la neuropsychologie. En effet, il semble bien que le système cognitif dissocie ces deux tâches (*cf.* Bruyer, 1986, Sergent & Signoret, 1992; Newcombe, Mehta, & de Haan, 1994; Farah, 1995).

La double dissociation entre agnosie des objets et prosopagnosie suggère cette séparation et de nombreuses dissociations expérimentales la confirme. Par exemple, il est difficile de choisir laquelle des photos inversées de la figure 1 représente la personne de la photo au-dessus, ou même si ces deux photos sont différentes. En les retournant la réponse devient évidente, ainsi que la transformation que l'image a subie. En revanche, le même genre de manipulation sur une photo d'objet comme le montre la figure 2 se détecte sans problème lorsque les photos sont inversées (pour une étude détaillée de ce phénomène, connu sous le nom d'illusion "Thatcher"[1], *cf.* Bartlett & Searcy 1993). Il convient toutefois de nuancer ce propos en rappelant que les tâches de la reconnaissance des objets diffèrent de celles de la reconnaissance des visages et que lorsque les tâches deviennent semblables, certaines différences s'estompent. Quoiqu'il en soit, nous évoquons tout d'abord les problèmes de la reconnaissance des objets, puis quelques développements récents de modélisation connexioniste de la reconnaissance des visages.

[1] Une preuve de plus de l'humour anglais, Thomson (1980), à qui l'on doit cette illusion, a, sans doute, pris plaisir à déformer le visage du premier ministre anglais de l'époque dont la popularité en milieu universitaire anglais était ce qu'elle était!

Une photo d'un visage!

A

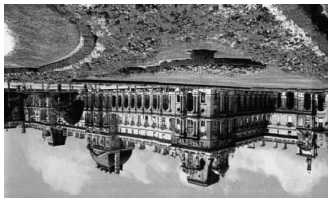
B

Laquelle de ces photos représente le visage du dessus?

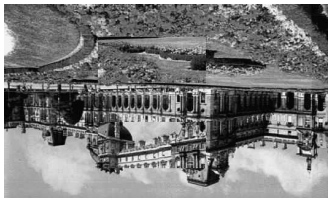
FIGURE 1. La fameuse illusion de Thatcher, d'après Thomson (1980).

Reconnaître un objet peut s'interpréter de plusieurs façons. J'*identifie* l'objet de la figure 3 comme étant une chaise. Autrement dit, l'objet appartient à la catégorie des chaises. Mais je

Une photo d'un monument célèbre.



A



B

Laquelle de ces photos représente le monument du dessus?

FIGURE 2. La fameuse illusion de Thatcher, n'est pas convaincante avec le Louvre!

peux aussi le reconnaître comme étant *ma chaise*. Pour l'essentiel, les modèles de la reconnaissance des objets ne sont concernés que par l'affectation des objets à leur catégorie. Bien entendu, il existe plusieurs façons de catégoriser un objet: est-ce une chaise, un objet de mobilier, une chaise de bureau, etc. (en fait ma chaise est une chaise spécifique: ces tâches représentent un continuum). En règle générale, les modèles cognitifs de la reconnaissance des objets visent à décrire la catégorisation au *niveau de base* comme le décrit Rosch (1978, *cf.* Abdi, 1986; Cordier, 1993; pour des revues de question). C'est à dire que l'objet de la figure 3 sera catégorisé comme une *chaise* plutôt qu'une *chaise de*

FIGURE 3. Un objet à reconnaître et à catégoriser.

bureau (niveau subordonné) ou qu'un *objet de mobilier* (niveau super-ordonné).

Pour reconnaître un objet se trouvant dans un espace à trois dimensions, le système visuel ne possède, en fait, qu'une information à deux dimensions fournie par l'oeil[2]. Par conséquent, une infinité d'objets en trois dimensions possèdent la même projection en deux dimensions (*cf.* les classiques illusions d'optique de la chambre d'Ames) et un même objet donne une infinité de projections différentes en deux dimensions. En toute rigueur, la reconnaissance des formes est un *problème mal posé*, car le système visuel ne possède pas suffisamment d'information pour

[2]La vision stéréoscopique n'est pas utilisée pour la reconnaissance des formes.

le résoudre. Néanmoins, nous reconnaissons facilement les objets qui nous entourent.

Selon Marr (1982), qui reste une influence théorique majeure, la reconnaissance des objets passe par la construction (ou le calcul) d'une vue de l'objet *indépendante* du point de vue de l'observateur appelée aussi *vue centrée sur l'objet*. Cette construction passe par plusieurs étapes de traitement de l'information visuelle comme par exemple la détection d'arrêtes illustrée par la figure 4. A partir des arrêtes, le système visuel extrait ensuite des formes élémentaires dont la combinaison permet alors d'identifier l'objet. Il existe plusieurs théories spécifiant les composantes élémentaires. (*e.g.,* Ullman, 1984; 1996; Pentland, 1986), parmi lesquelles la mieux connue est clairement celle proposée par Biederman (1987, 1995) sous le nom de théorie RBC (*Recognition By Components*) et plus connue sous le nom de théorie *géonique* de la reconnaissance des formes. Comme son nom l'indique, cette approche utilise comme composantes élémentaires des *géons*. Leur existence et identité précise est discutée plus bas.

1.1. *Un Géon, qu'est-ce?*

Comme nous l'avons vu, l'idée de base des modèles compositionels suppose que la reconnaissance des objets découle d'une décomposition des formes à reconnaître en constituants élémentaires. A ces formes élémentaires s'ajoute une grammaire. Ainsi un objet est uniquement déterminé par ses composants et leurs relations. Biederman (1987) appelle ces formes élémentaires des *géons* (un acronyme pour *geometrical ion*). Pour permettre une reconnaissance des objets qui soit indépendante du point de vue, les géons se doivent d'être eux-même des formes dont l'identification est possible sous le plus grand nombre de points de vue différents. Autrement dit, les géons sont des formes aussi *invariantes* que possible. Techniquement, ils sont créés à partir de la déformation d'un cylindre comme l'illustre la figure 5. L'idée de base provient, sans doute, de Marr (1977, *cf.* aussi Binford,

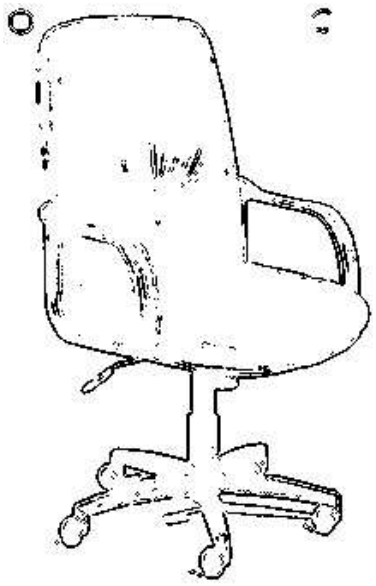


FIGURE 4. la chaise de la figure 3, filtrée par l'opérateur de Marr-Hidreth.

1981; Brook, 1981) et de son analyse indiquant que les *cônes généralisés* (une famille géométrique dont les géons sont un sous-groupe) peuvent servir de modèles pour décrire l'enveloppe du déplacement (continu) de formes convexes.

Selon Biederman, les géons existent en nombre limité (24 aux dernières nouvelles; Biederman, 1995, p. 143), et un petit nombre d'entre eux suffit pour identifier un objet à son niveau de base. La figure 6 montre quelques objets usuels et leur décomposition géonique. La figure 7 suggère une décomposition possible de la chaise de la figure 3. En comparant ces deux dernières figures, il est aussi apparent que la décomposition géonique ignore une partie importante de l'information initiale (en particulier la texture).

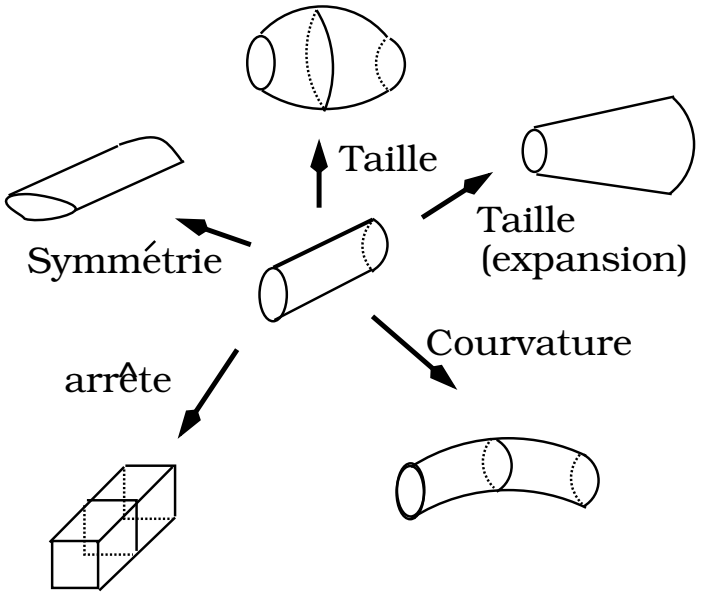


FIGURE 5. Comment créer des géons à partir d'un cylindre. D'après Biederman (1987).

1.2. *Support empirique pour les géons*

En plus de son élégance théorique toute structuraliste, l'approche géonique permet d'engendrer un bon nombre de prédictions testables (Hummel et Biederman, 1992, proposent également une instantiation connexioniste de la théorie, mais ne l'utilisent pas pour créer des prédictions spécifiques). Tout d'abord, comme la décomposition en géons ne demande que les arrêtes de l'image, la présence de l'information de texture ou même de couleur ne doit pas faciliter la reconnaissance d'objets usuels. Selon Biederman (1987) cette prédiction est vérifiée du moins dans des tâches de dénomination. En revanche, les modifications de l'image qui perturbent l'extraction de géons doivent également perturber l'identification des objets; alors que les modifications

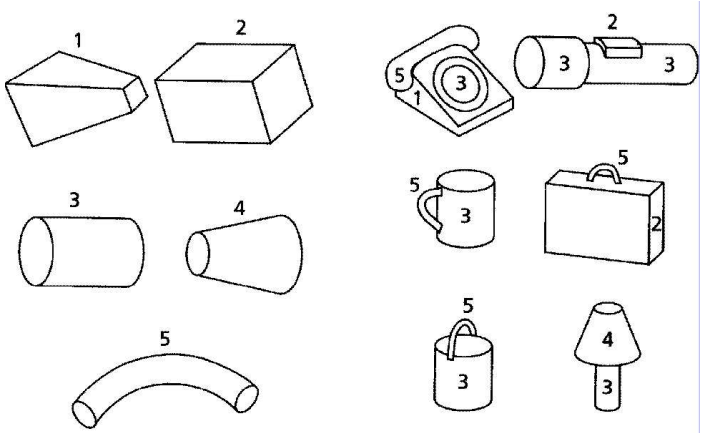


FIGURE 6. Quelques géons de base (à gauche) et quelques objets usuels (à droite) avec leur décomposition géonique selon Biederman (1990).

de l'image qui laissent l'identification géonique possible doivent avoir un effet mineur. Comme l'illustre l'exemple de la figure 8, cette prédiction semble se vérifier. Effectivement, lorsque Biederman (1987) demande à ses sujets de nommer des objets ainsi modifiés, le temps de réaction et la précision des réponses se trouvent affectés comme prévu.

1.3. *Problèmes avec les géons*

Malgré le nombre important de résultats expérimentaux en accord avec l'approche géonique, un certains nombre de résultats récents semblent difficiles à réconciler avec deux idées essentielles des géons: la notion de reconnaissance par composantes et la notion de construction d'un point de vue centré sur l'objet. Par exemple, Edelman et Bülthoff (1992, Bülthoff et Edelman, 1992; voir également Tarr, 1995), utilisent comme stimuli pour une épreuve de reconnaissance des objets qui ressemblent à des pièces

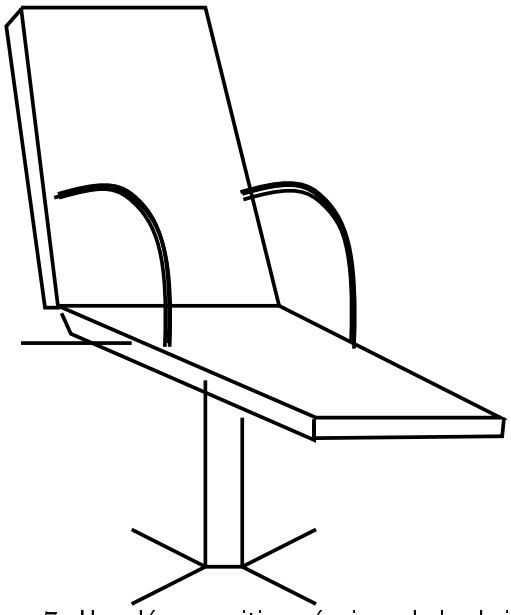


FIGURE 7. Une décomposition géonique de la chaise de la figure 3.

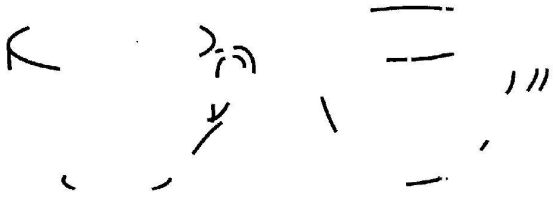


FIGURE 8. Une tasse dont la composition géonique est perturbée (à droite), ou n'est pas perburbée (à gauche. Il faut plus de temps et il est plus difficile d'identifier l'image de droite que celle de gauche.)

de fil de fer, à des "patatoïdes", ou à des amibes (*cf.* figure 9.) Les sujets apprenaient tout d'abord des vues de ces objets, puis

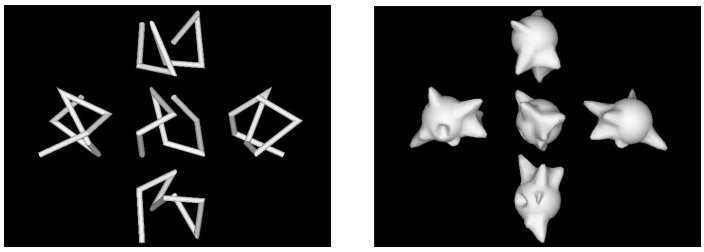


FIGURE 9. Quelques stimuli utilisés par Edelman & Bülthoff (1992).

devaient, dans une épreuve de choix forcé, reconnaître l'objet vu auparavant. Certaines nouvelles vues pouvaient se placer entre deux vues déjà apprises (ce qui correspond à une *interpolation*) ou se trouver extérieures aux vues apprises. Pour la théorie géonique, ces deux conditions sont équivalentes (puisque la composition géonique est invariante par rapport au point de vue). Les résultats obtenus ne s'accordent pas avec cette prédiction puisque les vues interpolées sont systématiquement mieux reconnues que les vues extérieures. Bülthoff et Edelman interprètent leurs résultats comme s'accordant avec un modèle proche du modèle du pandémonium de Selfridge (1959, voir aussi Lindsay et Norman, 1977). Pour eux la reconnaissance des formes s'obtient à partir d'une mesure de similarité entre la forme perçue et des formes apprises (sans qu'il y ait de décomposition géonique).

Ces résultats, bien que troublants, ne constituent pas, d'après Biederman (1995) une remise en cause fondamentale de la théorie géonique puisque, selon lui, les stimuli utilisés par Bülthoff et Edelman ne permettent pas une extraction facile de leur composition géonique. Il est difficile, dans l'état actuel, de trancher, mais la polémique entre ces différents auteurs (*i.e.,* Biederman *vs.* Tarr et Bülthoff et Edelman) est suffisamment intense pour qu'on puisse espérer des développements à rebondissements dans un futur proche.

Pour conclure, disons que, dans l'état actuel de la question, les théories componentielles (*e.g.,* la théorie géonique) permettent d'expliquer élégamment un grand nombre de phénomènes, mais qu'elles restent limitées à l'identification d'objets à leur niveau de base. Il n'est pas exclu que les différences de résultats (et d'interprétation) proviennent simplement d'une différence de *tâche* (*i.e.,* catégorisation pour les épreuves habituelles utilisées par Biederman opposée à l'identification d'un objet pour Tarr, Bülthoff et Edelman).

2. Des réseaux pour les visages

Si l'approche componentielle, et plus généralement les théories postulant la construction d'une vue centrée sur l'objet, permet de rendre compte d'un bon nombre de caractéristiques de la reconnaissance des objets, elle ne se montre pas si fructueuse avec les visages. Tout d'abord, du point de vue de la composition géonique, tous les visages sont équivalents et donc ne peuvent être distingués les uns des autres. Autrement dit, le problème essentiel de la reconnaissance des visages est *l'identification* (d'un objet semi-rigide parmi une classe dont les membres possèdent les mêmes caractéristiques).

Deux approches, essentiellement, abordent le problème de la reconnaissance des visages. La première suggère que les visages (de manière plus générale, les membres d'une même catégorie de base) peuvent se représenter par une ensemble de vues à deux dimensions (Lowe, 1987) dépendantes du point de vue de l'observateur. La reconnaissance d'un visage se ferait alors par comparaison avec les vues stockées en mémoire. Ce processus, équivalent à une reconnaissance des formes *par gabarit* (*cf.* Lindsay & Norman, 1977), soulève le problème habituel de ce type d'approche: comment stocker toutes ces vues, et comment prendre en compte les différences de taille perçues d'un même objet (*i.e.,* nous reconnaissons un visage qu'il soit vu de près ou de loin). Une solution possible est d'ajouter un processus

d'*interpolation* (comme vu plus haut pour les patatoïdes d'Edelman et Bültoff, 1992; *cf.* également Poggio & Edelman, 1990). Edelman & Weinshall (1991) montrent comment un réseau de neurones peut résoudre ce problème en utilisant des techniques d'interpolation non-linéaires.

Une solution plus économique en termes de stockage en mémoire revient à postuler qu'un visage est représenté par une seule ou un très petit nombre de *vues canoniques* ou de *vues prototypiques*[3] comme Palmer, Rosch et Chase (1981) l'ont suggéré pour la reconnaissance des objets. Les vues prototypiques révèlent l'information saliente pour la reconnaissance d'un objet et donc maximisent la reconnaissance. Ces vues sont systématiquement reconnues plus rapidement et plus précisement que les autres vues. Dans ce cadre, reconnaître un visage revient à transformer l'image perçue du visage pour pouvoir la comparer au prototype stocké en mémoire (*cf.* Tarr & Pinker, 1990). Cette approche ne demande qu'un nombre restreint d'images pour représenter un objet, mais suppose, en revanche, des processus de pré-traitement comme la normalization de la taille, la rotation mentale en trois dimensions, opérations gourmandes en calcul s'il en est. En outre, se pose le problème du choix et du développement des vues canoniques ainsi que de leur stockage et leur accès en mémoire à long terme.

2.1. *Deux codes pour les visages*

L'état actuel de la littérature ne permet pas de trancher entre les modèles évoqués plus haut. Néanmoins il semble qu'il y ait un changement dans la représentation des visages lorsqu'ils passent du statut de visages inconnus à celui de visages familliers (Bruce, 1982; Valentin, 1996). Ainsi, pour des visages inconnus, le changement de point de vue entre face et 3/4 face entre

[3]Le terme de *prototype* renvoie ici à la vue d'un objet la *meilleure* ou la *plus informative*, et non à une sorte de moyenne abstraite des différentes vues de l'objet.

l'apprentissage et le test diminue clairement la performance. La performance pour des visages familliers n'est pas affectée par de telles transformations ainsi que le confirment les anecdoctes où les changements de lunettes ou de coiffure des proches ne sont pas détectés (parfois à leur grand dam). Toutefois, même les visages familiers semblent sensibles au point de vue. Par exemple, Ellis (1986) montre que des sujets peuvent détecter après plus d'une semaine des changements de pose pour des visages familiers. Ceci suggère que deux types d'information sont stockés en mémoire correspondant à la distinction de Bruce et Young (1986) entre le code *pictorial* et le code *structural* ou à celle de Klatsky et Forest (1984) entre information *"vue-spécifique"* et *"visage-spécifique"*.

L'hypothèse d'un double codage s'accorde également avec les données de l'électrophysiologie et les mesures obtenues sur des neurones isolés. Par exemple, Perrett, Rolls et Caen (1982) rapportent l'existence, chez le singe rhésus, d'un groupe de neurones dans le *sulcus temporal* qui répondent sélectivement aux visages. Ces cellules répondent à des visages différents mais à un ensemble restreint de points de vue. Le passage d'un visage de face à profil réduit ou élimine la réponse de 60% des cellules, alors qu'une simple rotation de 10 ou 20 degrés entraîne une claire diminution des réponses. Par ailleurs, d'autres cellules répondent à un même visage quelque soit son orientation (*cf.* aussi Hasselmo, Rolls, Ballis & Nalwa, 1989, qui interprètent des données similaires comme une indication que les vues centrées sur l'objet sont construites à partir de différents points de vue). Selon Bruce (1982) l'information vue-spécifique et visage-spécifique est stockée pour les visages familliers, mais seule l'information vue-spécifique est gardée pour les visages inconnus.

La différence entre visages familliers et non-familliers pourrait s'expliquer par le plus grand nombre de points de vue stockés pour les visages familliers. Cette hypothèse s'accorde avec les résultats expérimentaux (Dukes & Bevan, 1968; Bartlett & Leslie,

1986) indiquant que des sujets voyant des visages non familliers sous différents points de vue deviennent moins sensibles à l'effet de rotation. Cet effet pourrait s'interpréter comme s'accordant avec le développement d'une représentation canonique tri-dimensionelle ou comme la résultante du stockage de différentes vues spécifiques.

Afin d'analyser plus en finesse ces interprétations, Valentin et Abdi (1996) simulent le stockage de différentes images bi-dimensionnelles de visages dans une mémoire auto-associative. Cette approche, déjà utilisée pour explorer la distinction entre code sémantique (*i.e.,* sexe, race) et code de l'identité (O'Toole, Abdi, Deffenbacher & Valentin, 1993), est équivalente à une analyse en composantes principales d'un ensemble d'images de visages (*cf.* Abdi, 1988, 1994a; Valentin, Abdi, O'Toole & Cottrel, 1994). Dans ce contexte, les composantes principales rebaptisées *vecteurs propres* (*eigenvectors* ou *eigenfaces*[4] en anglais) s'interprètent comme des constituants permettant de reconstruire des visages. A l'opposé, des caractéristiques locales comme le nez, la bouche, etc., les vecteur propres sont des "macro-caractéristiques" (Anderson & Mozer, 1981). Lorsque des images de visages prises avec différentes poses sont stockées dans un auto-associateur, les vecteurs propres (exprimant, donc, l'information extraite spontanément par le réseau de neurones) correspondent à deux types d'information perceptive différentes: l'un code la pose du visage (*i.e.,* face, profil), le second spécifie l'identité du visage. Cette dissociation est illustrée dans la figure 10 qui montre un visage reconstruit avec les premiers vecteurs propres (codant le point de vue), les vecteurs propres intermédiaires (dont le détail de la fonction reste à élucider) et les derniers vecteurs propres (codant l'identité).

L'examen des vecteurs propres eux-mêmes confirme l'existence de la dissociation entre vue et identité. La figure 11 montre les

[4]la traduction littérale par "visage propre" semble trop belle pour être vraie!

FIGURE 10. Une vue de face (en haut) et de profil (en bas) d'un visage et sa reconstruction à partir des vecteurs propres d'une mémoire auto-associative ayant appris 400 images de visages féminins (10 images par 40 visages). les colonnes 1 à 4 montrent respectivement la reconstruction avec 1) tous les vecteurs propres (ce sont les images originales) 2) les 20 premiers vecteurs propres 3) les vecteurs propres 21 à 100 et 4) les vecteurs propres 101 à 400. D'après Valentin (1996).

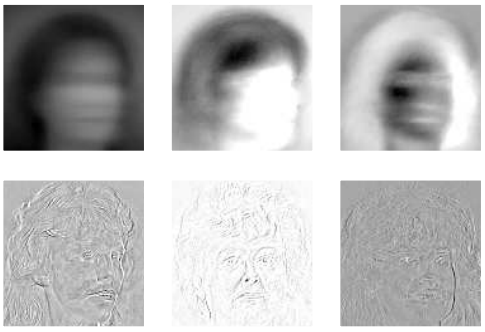


FIGURE 11. Les trois premiers vecteurs propres (en haut) et les trois derniers vecteurs propres d'une mémoire auto-associative ayant appris 10 images différentes de 40 visages féminins (soit 400 images en tout). D'après Valentin, 1996.

trois premiers et les trois derniers vecteurs propres de la mémoire auto-associative utilisée par Valentin et Abdi. Les trois premiers vecteurs propres semblent indiquer l'orientation du visage alors que les derniers vecteurs propres représentent l'identité d'un visage particulier sous un point de vue particulier. Une analyse plus statistique des données confirme cette intuition. Le premier vecteur propre représente une sorte de moyenne corrélée positivement avec toute image de visage, le second vecteur propre corrèle positivement avec les images de face et négativement avec les images de profil. Il apparaît ainsi comme un détecteur de position. Cette interprétation se confirme dans la figure 12 où l'addition des deux vecteurs propres donne une vue de profil, alors que soustraire le deuxième vecteur propre du premier donne une vue de face. Cette dissociation entre orientation et identité montrée par la mémoire auto-associative rappelle la dissociation observée par Perrett *et al.* pour les neurones du singe rhésus.

2.2. *Reconnaissance de nouvelles vues d'un visage*

Les simulations précédentes montrent une dissociation entre identité et orientation, mais le problème de la reconnaissance de nouvelles vues d'un visage reste entier. Pour y répondre, Valentin et Abdi (1996) dans une nouvelle série de simulations font apprendre à une mémoire auto-associative 1, 2 ou 9 vues de 15 visages différents. La mémoire essaie alors de reconstituer (*i.e.*, de reconnaître) de nouvelles vues des visages appris et de nouvelles vues de visages inconnus. La qualité de la reconstruction, mesurée par la corrélation entre l'image reconstituée et l'original, est considérée comme analogue à une mesure de familiarité et permet de construire des courbes ROC qui expriment la performance de la mémoire en termes de la théorie de la détection du signal. Comme le montre la figure 13, plus grand est le nombre de vues différentes d'un même visage, meilleure est la performance de la mémoire. La performance de la mémoire après apprentissage d'une seule vue par visage est équivalente au hasard, elle

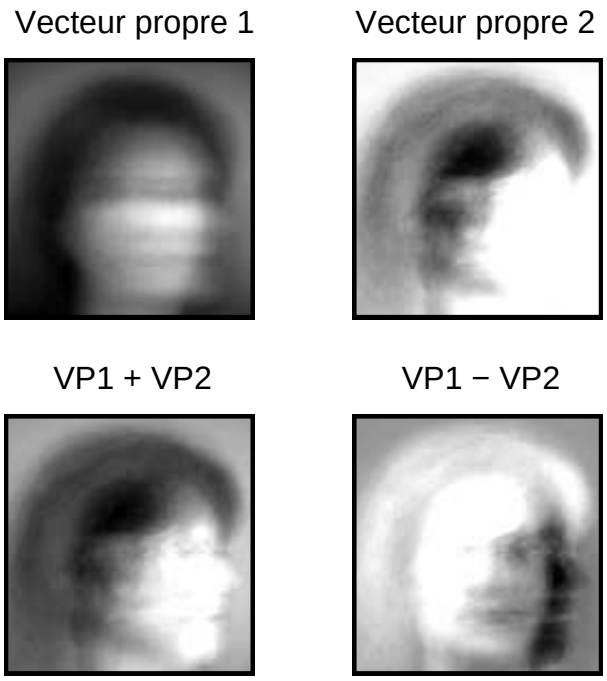


FIGURE 12. Illustration du rôle de détecteur de position du premier et du deuxième vecteur propre d'une mémoire auto-associative: 1) Le premier vecteur propre (en haut à gauche); 2) le deuxième vecteur propre (en haut à droite); 3) le premier vecteur propre plus le second vecteur propre (en bas à gauche) donne une vue de profil; 4) le premier vecteur propre moins le second vecteur propre (en bas à droite) donne une vue de face. D'après Valentin & Abdi (1996).

n'est pas très impressionnante avec 4 vues différentes, mais elle le devient avec 9 vues différentes.

L'analyse des résultats en function de la vue utilisée pour tester la performance montre que certaines vues d'un visage sont

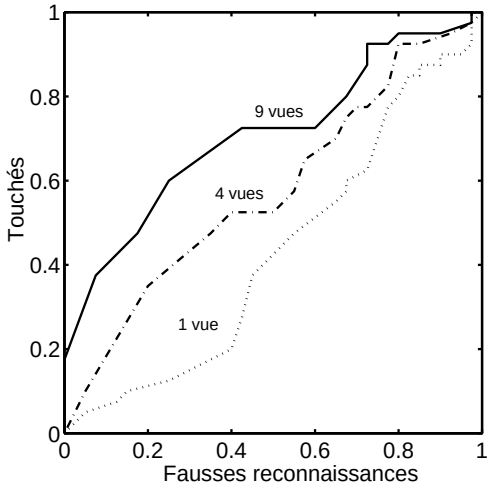


FIGURE 13. Courbes ROC en fonction du nombre de vues apprises. La surface sous la courbe donne une estimation non-biaisée du pourcentage d'identifications correctes. Une performance aléatoire (*i.e.,* 50%) correspond à la diagonale allant du coin en bas à gauche au coin en haut à droite. la condition 1 vue ne fait pas mieux que le hasard, 4 vues un peu mieux, et 9 vues clairement mieux.

plus faciles à reconnaître[5] que d'autres (*cf.* la figure 14). En particulier, la vue de 3/4 face conduit systématiquement à une meilleure performance que les autres vues. Autrement dit, la vue de 3/4 est *la plus facile* à reconnaître. Ce phénomène vaut la peine d'être souligné puisqu'il correspond à "l'avantage du 3/4 face" observé chez les sujets humains (Bruce, Valentine, & Baddeley, 1987; Bruyer & Galvez, 1989; Fagan, 1979; Krouse, 1981; Logie, Baddeley, & Woodhead, 1987; Valentin 1996). Cet effet est parfois interprété comme une indication que la vue de 3/4

[5]pour parodier un mot célèbre: certaines vues sont plus égales que d'autres!

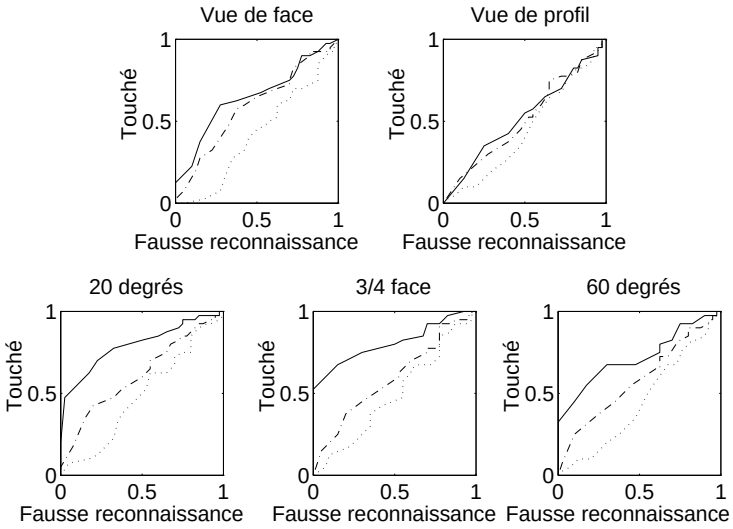


FIGURE 14. Courbes ROC en fonction du nombre de vues apprises et de la vue au test. La surface sous la courbe donne une estimation non-biaisée du pourcentage d'identifications correctes. Une performance aléatoire (*i.e.,* 50%) correspond à la diagonale allant du coin en bas à gauche au coin en haut à droite. En trait continu: condition 9 vues, en trait mixte fin: condition 4 vues, en pointillé: condition 1 vue. La vue de 3/4 correspond à la meilleure performance.

constitue le prototype ou la vue canonique évoqués plus haut (Palmer *et al.*, 1981).

Ainsi que le notent Valentin, Abdi et Edelman (1996), cette interprétation de la vue de 3/4 face comme vue canonique paraît en désaccord avec les résultats de l'électrophysiologie. En effet, les cellules sensibles au point de vue trouvées chez le singe montrent un pic de sensitivité pour la vue de profil ou pour la vue frontale (parfois aussi pour l'arrière de la tête) mais pas pour le 3/4 face (Desimone, Albright, Gross, & Bruce; 1984; Perrett *et al.*, 1986

Perrett, Mistin, & Chitty, 1987; Perrett, Smith, Potter, *et al.*, 1985). Perrett *et al.*, interprètent ces résultats comme indiquant que "la reconnaissance de chaque individu connu de l'observateur s'effectue par l'analyse d'un petit ensemble de vues prototypiques de cet individu (p. 191)". Perrett *et al.*, suggèrent que les vues intermédiaires entre la vue de face et de profil sont reconnues par interpolation entre ces vues prototypiques ou canoniques. En particulier, la vue de 3/4, en l'absence de cellule spécifique, devrait activer également les cellules spécifiques de la vue de face et de la vue de profil.

Comme il se doit, toutefois, le fait de ne pas trouver de cellules spécifiques du 3/4 ne signifient pas qu'elles n'existent pas. En fait, des travaux plus récents semblent indiquer l'existence de quelques cellules sensibles à d'autres vues que face ou profil (Hasselmo *et al.*, 1989; Perrett *et al.*, 1989, 1991). Neanmoins après une ré-analyse des résultats, Perrett *et al.* (1994) concluent que

> des études qualitatives et quantitatives récentes ont, toutefois, confirmées la notion de codage préférentiel de certaines vues. Bien que les cellules soient accordées (tuned) à un ensemble large de point de vues, il y a une préférence statistique claire pour la vue de profil et de face (p.50–1).

Ainsi les résultats de l'électrophysiologie s'accordent avec l'hypothèse originale de Perrett *et al.* considérant les vues de face et de profil comme des vues canoniques. En accord avec cette hypothèse, Harries, Perrett, et Lavender (1991) rapportent que des sujets passent plus de temps à examiner des vues de face et de profil que toute autre vue lorsqu'ils apprennent des visages.

Pour résumer: les données de la psychologie expérimentale (essentiellement performance en termes de reconnaissance) suggèrent que la vue de 3/4 face joue le rôle de prototype, alors que les données en provenance de l'électrophysiologie s'accordent avec l'hypothèse de l'existence de deux vues canoniques: face et profil. En fait, Valentin *et al.*, démontrent que cette contradiction n'est

qu'apparente. Pour ce faire, ils utilisent un réseau de neurones spécifique appelé réseau à fonction de rayon (*radial basis function*, *cf.* Abdi, 1994b; pour plus de détails) abrégé en "réseau RBF". Pour l'essentiel, ces réseaux simulent des prototypes. Ils sont composés d'une couche de cellules d'entrée (qui joue le rôle d'une rétine), d'une couche de cellules cachées qui jouent le rôle de prototypes et d'une couche de cellules externes qui jouent le rôle de détecteurs d'identité. Les cellules de la couche cachée calculent la similarité[6] entre le prototype et l'image présentée sur la rétine. Puis, elles transmettent cette similarité aux cellules de la couche externe qui adaptent la valeur de leur connexions de façon à maximiser la reconnaissance. L'avantage essentiel de ces réseaux est de pouvoir choisir les prototypes et leur représentation. La figure 15 détaille le schéma d'une telle architecture. Valentin *et al.* utilisent la propriété des réseaux RBF de représenter explicitement les prototypes par les cellules de la couche cachée et construisent cinq réseaux différents chacun correspondant à un type de représentation interne ou à une hypothèse de représentation spécifique. Une brève description de chacun d'entre eux suit:

- L'hypothèse selon laquelle toutes les vues sont stockées est implémentée en affectant une cellule de la couche cachée à chaque vue apprise. On peut s'attendre à ce que cette condition soit optimale puisque toute l'information apprise est gardée. C'est le modèle à exemplaires ultime.

[6] pour les amateurs d'émotions fortes et de belles formules: la similarité est calculée comme une fonction gaussienne de la distance euclidienne. Si d est la distance euclidienne entre le prototype et l'objet présenté, la similarité notée s sera:

$$s = \exp\left\{-\frac{d^2}{2\sigma^2}\right\}$$

où σ est un paramètre (parfois libre) de dispersion (*i.e.*, l'écart type de la distribution Gaussienne).

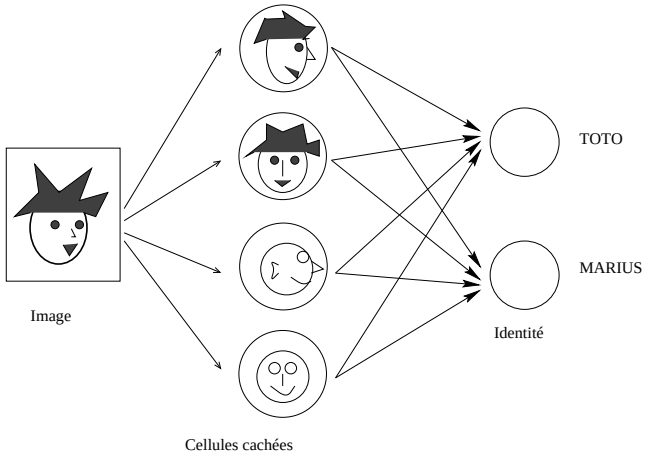


FIGURE 15. Schéma d'un réseau RBF. L'image présentée sur la rétine est transmise aux cellules de la couche cachée qui jouent le rôle de prototypes. Chacune d'entre elles calcule la similarité entre le prototype qu'elle a stocké et l'image présentée. Les cellules de la couche cachée transmettent ensuite leur activation aux cellules de la couche de sortie qui ajustent la valeur de leur connexion de façon à maximiser l'identification de l'image.

- L'hypothèse d'abstraction d'une vue unique moyenne (ou d'un prototype central) représente chaque visage par la moyenne des vues apprises. En fonction de la parabole des "choux et des carottes" on s'attend à ce que cette condition conduise à la pire performance.
- L'hypothèse d'une vue canonique conduit à représenter chaque visage par sa vue de 3/4.
- l'hypothèse de deux vues canoniques représente chaque visage par une vue de profil et de face.
- l'hypothèse la plus généreuse de vues canoniques représente chaque visage par trois vues: face, 3/4 et profil.

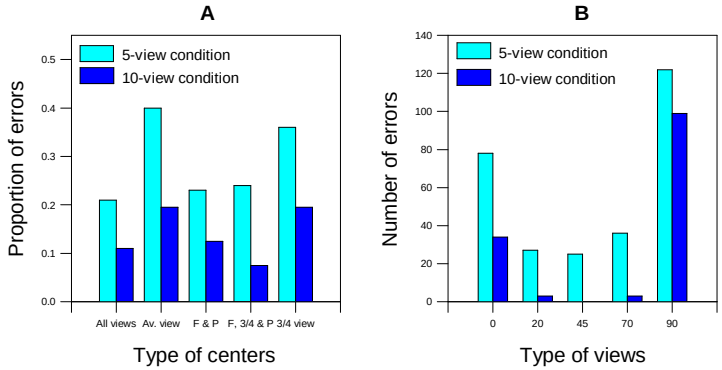


FIGURE 16. Proportion d'erreur de reconnaissance comme une fonction du type de prototype et du nombre de vues apprises (à gauche) et du type de vue testées (à droite). F désigne une vue Frontale et P une vue de Profil. D'après Valentin (1996).

Chaque réseau apprend 4 vues (cette condition est appelée condition 5 vues puisque 5 vues de chaque visage [4 appris+ 1 testé] sont utilisées) ou 9 vues (condition 10 vues) de chaque visage puis est ensuite testé sur sa capacité à reconnaître un visage appris à partir d'une nouvelle vue. Les auteurs rapportent leur résultats en utilisant la proportion d'erreur d'identification.

Comme on peut le voir dans la figure 16, il n'est pas nécessaire de créer une représentation interne invariante pour *identifier* de nouvelles vues de visages connus. Il n'est pas nécessaire, non plus, de mémoriser toutes les vues. En fait, deux vues (face et profil) suffisent pour obtenir 90% d'identifications correctes. En termes de distance à une vue prototypique ou canonique, ces résultats s'accordent avec les mesures de neurones isolés: deux vues canoniques (face et profil) sont supérieures à une seule (le 3/4). En outre, comme l'ajout de la vue de 3/4 n'améliore pas la performance du système, on peut supposer que les vues de face et de profil sont optimales pour la reconnaissance. Enfin,

la supériorité de la vue de 3/4 (ou plus généralement des vues intermédiaires) se retrouve quelque soit le type de représentation utilisée.

Que ressort-il de ces ensembles de simulations? Peut-être avant tout une leçon de prudence: Il ne suffit pas d'observer une performance supérieure pour une vue particulière pour en inférer la prototypie. Pour les visages, ces résultats indiquent qu'un nombre important de résultats peuvent s'obtenir sans qu'il soit nécessaire de présupposer une extraction d'invariant de formes. En fait, la condition la plus importante pour les modèles connexionistes actuels de la reconnaissance des visages est d'utiliser des images de visages réels (plutôt que des vecteurs aléatoire ou arbitraires) et des opérations simples de pré-traitement (*e.g.*, normalisation de taille, filtrage).

3. En conclusion

Il y a quelques années, la reconnaissance des objets semblait nécessairement passer par l'abstraction d'un code ou d'une représentation de l'objet indépendante du point de vue de l'observateur. Après quelques années de modélisation et d'expérimentation il apparaît, maintenant, que les visages échappent à cette règle et que des modèles spécifiques aux visages à base de représentations en 2D peuvent expliquer les données expérimentales de manière plus satisfaisante que les modèles classiques de la reconnaissance des objets. Il reste à voir si, en retour, les attaques récentes envers les modèles de reconnaissance des objets supposant une décomposition des objets en géons (ou autres éléments de base), vont conduire à une ré-évaluation essentielle des pré-supposés théoriques de la reconnaissance des objets: *"Stay Tuned"* pourrait-on dire. La suite risque forte d'être pleine de surprises.

Bibliographie

[1] Abdi, H. (1986) La mémoire sémantique une fille de l'intelligence artificielle et de la psychologie. In C. Bonnet, J.M. Hoc, G. Tiberghien

(Eds.) *Psychologie, intelligence artificielle et automatisme*. pages 139–151. Mardaga, Bruxelles.

[2] Abdi, H. (1988). Generalized approaches for connectionist auto-associative memories: Interpretation, implication, and illustration for face processing. In J., Demongeot, editor, *Artificial intelligence and cognitive sciences*, pages 151–164. Manchester University Press, Manchester.

[3] Abdi, H. (1994a). *Les réseaux de neurones*. Presses Universitaires de Grenoble, Grenoble.

[4] Abdi, H. (1994b). A neural network primer. *Journal of Biological Systems*, 2:247–281.

[5] Anderson, J. & Mozer, M. (1981). Categorization and selective neurons. In Hinton, G. & Anderson, J., editors, *Parallel models of associative memory*, pages 213–236. Erlbaum, Hillsdale.

[6] Bartlett, J. & Leslie, J. (1986). Aging and memory for faces versus single views of faces. *Memory and Cognition*, 14:371–381.

[7] Bartlett, J. C. & Searcy, J. (1993). Inversion and configuration of faces. *Cognitive Psychology*, 25:281–316.

[8] Biederman, I. (1987). Recognition by components: A theory of human image understanding. *Psychological Review*, 94:115–145.

[9] Biederman, I. (1990). Higher-level vision In D.N. Osherson, S.M. Kosslyn, J.M Hollerbach (Eds.) *Visual cognition and action*. pages 41–72. MIT Press, Cambridge.

[10] Biederman, I. (1995). Visual object recognition In S.M. Kosslyn, D.N. Osherson (Eds.) *Visual cognition*. pages 121–165. MIT Press, Cambridge.

[11] Binford, T.O. (1981). Inferring surfaces from images. *Artificial Intelligence*, 17:205-244.

[12] Brooks, R. (1981). Symbolic reasoning among 3-dimensional models and 2-dimensional images. *Artificial Intelligence*, 17:285–349.

[13] Bruce, V. (1982). Changing faces: Visual and non-visual coding process in face recognition. *British Journal of Psychology*, 73:105–116.

[14] Bruce, V., Valentine, T., & Baddeley, A. (1987). The basis of the 3/4 advantage in face recognition. *Applied Cognitive Psychology*, 1:109–120.

[15] Bruce, V. & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, 77:363–383.

[16] Bruyer, R. (1986). *The neuropsychology of face perception and facial expression*. Lawrence Erlbaum, Hillsdale.

[17] Bruyer, R. & Galvez, C. (1989). The structural orientation of the mental representation of faces. *Archives de Psychologie*, 57:259–269.

[18] Bülthoff, H. & Edelman, S. (1992). Psychological support for a two dimensional view interpolation theory of object recognition. *Proceeding of the National Academy of Science*, 89:60–64.

[19] Cordier, F. (1993). *Les représentations cognitives privilégiées: Typicalité et niveau de base*. Presses Universitaires de Lille, Lille.

[20] Desimone, R., Albright, T., Gross, C., & Bruce, C. (1984). Stimulus selective properties of inferior temporal neurons in macaque. *Journal of Neuroscience*, 8:2051–2062.

[21] Dukes, W. & Bevan, W. (1968). Stimulus variation and repetition in the acquisition of naming responses. *Journal of Experimental Psychology*, 74:178–181.

[22] Edelman, S. & Bülthoff, H. (1992). Orientation dependence in the recognition of familiar and novel views of three dimensional objects. *Vision Research*, 32:2385–2400.

[23] Edelman, S. & Weinshall, D. (1991). A self-organizing multiple-view representation of 3d objects. *Biological Cybernetics*, 64:209–219.

[24] Ellis, H. (1986). Introduction: Process underlying face recognition. In Bruyer, R., editor, *The neuropsychology of face perception and facial expression*. Erlbaum, Hillsdale.

[25] Fagan, J. (1979). The origins of facial pattern recognition. In Bornstein, M. & Keesen, W., editors, *Psychological development from infancy: Image to intention*. Erlbaum, Hillsdale.

[26] Farah, M. J. (1995). Dissociable systems for visual recognition: A cognitive neuropsychological approach. In Kosslyn, S. M. & Osherson, D. N., editors, *Visual Cognition*, Cambridge. MIT press.

[27] Harries, M., Perrett, D., & Lavander, A. (1991). Visual inspection during encoding and recognition of 3d heads. *Perception*, 20:669–680.

[28] Hasselmo, M., Rolls, E., Baylis, G., & Nalwa, V. (1989). Object-centered encoding by face-selective neurons in the cortex in the superior temporal sulcus of the monkey. *Experimental Brain Research*, 79:417–429.

[29] Hummel, J. E. & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99:480–517.

[30] Klasky, R. & Forrest, F. (1984). Recognizing familiar and unfamiliar faces. *Memory and Cognition*, 12:60–70.

[31] Krouse, F. (1981). Effects of pose, pose change, and delay on face recognition performance. *Journal of Applied Psychology*, 66:651–654.

[32] Lindsay, P.H., Norman, D.A. (1977). *Human information processing*. Academic Press, New York.

[33] Logie, R., Baddeley, A., & Woodhead, M. (1987). Face recognition, pose and ecological validity. *Applied Cognitive Psychology*, 1:53–69.

[34] Lowe, D. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:335–395.

[35] Marr, D. (1977). Analysis of occluding contours. *Proceedings of the Royal Society of London*, B197: 441-475.

[36] Marr, D. (1982). *Vision*. Freeman: San Francisco.

[37] Newcombe, F., Mehta, Z., de Haan, E.H.F. (1994) Issues of representation in object vision. In Farah, M. & Ratcliff, G., editors, *The neuropsychology of high level vision: Collected tutorial essay*, pages 33–61. Erlbaum, Oxford.

[38] O'Toole, A., Abdi, H., Deffenbacher, K., & Valentin, D. (1993). A low dimensional represention of faces in the higher dimensions of the space. *Journal of the Optical Society of America A*, 10:405–411.

[39] Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In Long, J. & Baddeley, A., editors, *Attention and performance IX*. pages 135–151. Erlbaum, Hillsdale.

[40] Pentland, A. (1986). Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28:293–331.

[41] Perrett, D., Mistlin, A., & Harries, M. (1989). Seeing faces: The representation of facial information in temporal cortex. In Kulikowski, J., Dickinson, C., & Murray, I., editors, *Seeing contour and colour*, pages 770–754. Pergamon, Oxford.

[42] Perrett, D., Mistlin, A., Potter, D., Smith, P., Head, A., Chitty, A., Broennimann, R., Milner, A., & Ellis, M. (1986). Functional organization of visual neurons processing face identity. In Ellis, H., Jeeves, M., Newcombe, F., & Young, A., editors, *Aspects of face processing*, pages 187–198. Nijhoff, Dordrecht.

[43] Perrett, D., Mistlin, J., & Chitty, A. (1987). visual neurons responsive to faces. *Trends in Neurosciences*, 10:358–364.

[44] Perrett, D., Oram, M., Harries, M., Bevan, R., Hietanen, J., Benson, P., & Thomas, S. (1991). Viewer-centered and object-centered coding of heads in the macaque temporal cortex. *Experimental Brain Research*, 86:159–173.

[45] Perrett, D., Oram, M., Hietanen, J., & Benson, P. (1994). Issues of representation in object vision. In Farah, M. & Ratcliff, G., editors, *The neuropsychology of high level vision: Collected tutorial essay*, pages 33–61. Erlbaum, Oxford.

[46] Perrett, D., Rolls, E., & Caan, W. (1982). Visual neurons responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, 47:329–342.

[47] Perrett, D., Smith, P., Potter, D., Mistlin, A., Head, A., Milner, A., & Jeeves, M. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London B*, 223:293–317.

[48] Poggio, T. & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266.

[49] Rosch, E. (1978). Principle of categorization. In Rosch, R. & Lloyd, B., editors, *Cognition and categorization*, pages 27–48. Erlbaum, Hillsdale.

[50] Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

[51] Selfridge, O.G. (1959). Pandemonium: A pardigm for learning. In *The mechanisation of thought processes*. H.M. Stationary office, London.

[52] Sergent, J. & Signoret, J. (1992). Functional and anatomical decomposition of face processing: Evidence from prosopagnosia and study of normal subjects. *Philosophical Transaction of the Royal Society of London (B)*, 335:55–62.

[53] Tarr, M. & Pinker, S. (1990). Mental rotation and orientation dependence in shape recognition. *Cognitive Psychology*, 21:233–282.

[54] Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review*, 2:55–82.

[55] Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, 9:483–484.

[56] Ullman, S. (1984). Visual routines. *Cognition*, 18:97–159.

[57] Ullman, S. (1989). Aligning pictorial description: An approach to object recognition. *Cognition*, 32:193–254.

[58] Ullman, S. (1996). *High-level vision*. MIT Press, Cambridge.

[59] Valentin, D. (1996). *How come when you turn your head I still know who you are; Evidence from computational simulations and human behavior*. PhD thesis, University of Texas at Dallas.

[60] Valentin, D. & Abdi, H. (1996). Can a linear autoassociator recognize faces from new orientations? *Journal of the Optical Society of America A*, 13:717–724.

[61] Valentin, D., Abdi, H., & Edelman, B. (1995, November). *Recognizing faces from new view angles: Human subjects and computational model evidence*. Paper presented at the 36th Annual Meeting of the Psychonomic society, Los Angeles, CA.

[62] Valentin, D., Abdi, H., O'Toole, A., & Cottrell G. (1994). Connectionist model of face processing: A survey. *Pattern Recognition*, 27:1209–1230.