ETUDE DES CRITERES DE CATEGORISATION D'IMAGES DE SCENES NATURELLES COMPLEXES

Clément METGE, Alain PUGOL, Mohsen ARDABILIAN

LIRIS, Ecole Centrale de Lyon,
CNRS-UMR 5205

Mai 2009

Résumé :

Cette étude s'inscrit dans une démarche d'enrichissement mutuel entre les sciences de l'ingénieur et les sciences cognitives. La question initialement posée est "comment catégorise-t-on une image de scène naturelle complexe ?" Après une sous spécification de la question, afin de la préciser, puis une étude bibliographique, cette étude propose un paradigme expérimental original. Nous présentons aux sujets des images de villes et de forêts. Nous faisons varier la proportion des éléments, le centrage ainsi que le filtrage des images. Nous mesurons la classification des sujets et le temps de réponse. Les sujets sont répartis en deux groupes, un groupe ayant deux choix et un groupe ayant trois choix catégoriels. Les résultats montrent que la proportion, le centrage, et le filtrage exercent une influence significative sur la catégorisation; la proportion exerce une influence significative sur les temps de réponse.

Mots clés :

Images naturelles complexes. Classifieur. Catégorisation. Catégories. Proportion. Centrage. Filtrage passe bas. Filtrage passe haut. Traitement de l'image.

Table des matières

1 Introduction

Cette recherche s'inscrit dans un axe de recherche en intelligence artificielle, dont le but est de simuler l'activité humaine de catégorisation d'images naturelles complexes, afin de mettre au point un classifieur automatique d'image.

La question initialement posée était "Comment catégorise-t-on une image de scènes naturelles complexes ?". Nous avons réduit la question en "Selon quels critères une image fait partie d'une catégorie, et pas d'une autre ?". Nous avons donc cherché à isoler certains de ces critères perceptifs globaux. Après un état de l'art sur l'activité de catégorisation, sur les différentes techniques de traitement du signal, et sur les protocoles expérimentaux couramment utilisés pour étudier cette tâche, nous avons décidé la création d'un nouveau protocole pour répondre à notre problématique. Après avoir créer les images à partir de photos réelles, et avoir implémenter le programme. Nous présentons aux sujets des images de villes et de forêts. Nous faisons varier la proportion des éléments, le centrage ainsi que le filtrage des images. Nous mesurons la classification des sujets, ainsi que les temps de réponse. Les sujets sont répartis en deux groupes, un groupe ayant deux choix de réponse catégorielle, et un groupe ayant trois choix. Nous avons fait passer l'expérience à 40 sujets. Les résultats nous montrent que la proportion, le centrage et le filtrage ont un impact significatif sur l'activité de catégorisation. La proportion joue également un rôle significatif sur les temps de réponse. Les résultats de cette étude serviront de base pour établir un classifieur d'image par analyse de contenu.

1.1 Buts et applications de la recherche

L'étude s'inscrit dans une problématique d'accès automatique à de très grandes bases d'images. Comment trouver dans une base de plusieurs milliers d'images, celles correspondant au mieux à une description succincte d'une scène ? Ainsi un des buts de cette recherche est de permettre la mise au point d'un moteur de recherche, un classifieur automatique capable de catégoriser automatiquement des images de tous types, par analyse de contenu. Nous avons, pour le moment, restreint les images aux scènes naturelles complexes, sachant que tous types d'images doit être pris en compte, vu l'énorme hétérogénéité que l'on peut trouver sur la toile. Un classifieur peut catégoriser de multiples manières, en fonction de très nombreux critères, présences de couleurs, de lignes horizontales, ou verticales, répartitions des fréquences… Le but avoué de la recherche est que le moteur classifie avec des critères qui soient pertinents pour l'homme. Il s'agit donc d'isoler ces critères, et de les implémenter dans le robot. Nous

avons choisis de tester, dans cette étude, les critères de proportion, de centrage, et de filtrage, qui nous paraissent pertinents.

2 Eléments de bibliographie

2.1 Perception et catégorisation humaine

2.1.1 Modèles de perception visuelle

Il existe de nombreuses approches concernant l'étude de la perception visuelle. Ce domaine concerne des chercheurs aussi variés que des psychologues, des biologistes, des neurobiologistes, des ingénieurs, des informaticiens ou des mathématiciens.

Nous pouvons dégager trois familles :

- L'approche psycho-visuelle, rattachée aux aspects psychologiques de la perception visuelle.
- L'approche analytique, qui cherche à comprendre le fonctionnement des mécanismes sensoriels et neuronaux de la vision au niveau biologique.
- L'approche calculatoire, qui traite des problèmes algorithmiques de l'acquisition, du traitement et de l'interprétation des informations visuelles.

Nous allons nous attarder sur l'approche calculatoire de la perception visuelle. Celle-ci a progressivement évolué en deux approches principales. La plus ancienne, l'approche de vision reconstructive correspond à la conception empiriste de la vision et dérive essentiellement des travaux de D.Marr. La plus récente est celle de vision intentionnelle. Elle hérite d'une conception dynamique et active de la vision liée aux travaux de Gibson.

2.1.2 Le paradigme reconstructif

Le paradigme reconstructif (Recovery Paradigm) est l'approche classique de la vision par ordinateur. Elle consiste à tirer du monde visible une représentation symbolique de ses propriétés, à la fois géométrique et physique, et à exploiter cette représentation pour un certains nombre de tâche de haut niveau : reconnaissance, localisation, catégorisation, etc. Héritière de l'école Empiriste de la perception visuelle, cette approche part du principe que le monde possède une structure, et par conséquent un certain nombre de régularités qui sont utiles à sa représentation.

2.1.3 Le paradigme de Marr et les processus descriptif

Entre l'intelligence artificielle, les théories de l'information, la cybernétique, l'informatique, David Marr propose le premier cadre d'une approche de la vision d'un point de vue

calculatoire. Son travail, fondamental pour la vision par ordinateur, est aussi une tentative de clarifier la conception des systèmes de traitement de l'information. Selon lui, le but de la vision en tant que système de traitement de l'information est de décrire l'environnement extérieur. De la rétine à la conscience du monde, l'activité du cerveau s'applique à une représentation du monde qui nous est intérieur. Les neurones ne manipulent pas d'images mais une représentation symbolique d'une scène élaborée à partir d'images. La vision est rapportée au problème de la construction de cette représentation. Marr propose une méthodologie pour l'analyse de ces processus. Trois niveaux sont ainsi définis :

- Le niveau calculatoire : Ce niveau cherche à définir les buts du système.
- Représentation et algorithme : Ce niveau concerne la modélisation du système proprement dit, comment passer d'une étape à l'autre. Quels sont les algorithmes nécessaires ? Quelle est la structure de représentation ?
- Implémentation physique : C'est le niveau matériel du système. Comment implémenter, quels sont les capteurs les plus pertinents ? Quel langage utiliser, et quel type de support ?

Dans ce contexte, la vision fonctionne comme un système de traitement de l'information, en tant que tel, il est nécessaire de la diviser en niveau de traitements. Marr adopte une approche modulaire pour simplifier ce problème complexe en une hiérarchie de problèmes plus simples. Ainsi il fragmente le problème en différents modules de perception. Il propose une décomposition en quatre niveaux de traitements :

- L'image : C'est le niveau le plus bas, celle de l'image rétinienne. Sa fonction est de représenter la distribution de l'intensité lumineuse sur la rétine.
  L'ébauche primaire : "primal sketch" : processus de bas niveau :
  La fonction de ce niveau est de rendre explicite, à partir des intensités lumineuses, des informations géométriques sur la façon dont elles sont organisées. La possibilité de détecter des surfaces commence à ce niveau.
- L'ébauche 2D1/2 : Processus de niveau intermédiaire :
  Ce niveau rend explicite l'orientation des surfaces, fournit une estimation de la profondeur, de la proportion, ainsi que toute information sur l'organisation de l'image. A ce niveau la représentation n'a pas encore un caractère global.
- Représentation 3D : Processus de haut niveau :
  Ce niveau donne une représentation tridimensionnelle des objets indépendamment de l'observateur. C'est une représentation symbolique de la scène telle qu'elle est perçue. C'est à ce niveau qu'il y a extraction de sens.

Dans cette étude nous nous intéressons essentiellement aux processus de niveaux intermédiaires.

2.1.4 L'activité de catégorisation chez l'homme

Chez l'homme, la reconnaissance visuelle de scènes complexes est généralement rapide, automatique et fiable (Thorpe, 2001). Cette apparente simplicité contraste avec la difficulté à décrire et modéliser les traitements mis en jeu dans le processus de reconnaissance, identification visuelle et catégorisation.

Il est important de noter que nous parlons ici de "catégorie perceptive" et non de "catégories sémantiques". La sémantique décrit la nature et la fonction des objets. Le niveau perceptif correspond à la forme et à la géométrie. Ce que nous cherchons ici ce sont les caractéristiques communes aux images d'une même catégorie, sans prendre en compte des caractéristiques de plus haut niveau. Les systèmes artificiels décrivent les images à partir d'attributs dits de bas niveau (couleur, texture, distribution d'orientation, relation spatiale…) Alors que les sujets humains décrivent l'image en utilisant des conceptions plusieurs niveaux. Il faut donc chercher à savoir quels sont les attributs utilisés et quels sont les primordiaux. On peut considérer que de nombreux problèmes de perception se ramènent d'une manière ou d'une autre à un problème de catégorisation. Il est donc important de déterminer la notion de catégorie. Dans un dictionnaire, on trouve cette définition de *catégorisation* ou *catégorie*.

Catégorie: (lat. categoria; gr. χατηγορια, attribut, de χατα, sur, et αγορευειν, parler). Dans le langage ordinaire, toute classe dans laquelle on range plusieurs objets présentant des caractères communs.

Jusqu'à la fin des années 1970, le cadre d'analyse des catégories et des processus de catégorisation s'inscrit dans une conception aristotélicienne, s'appuyant notamment sur les travaux pionniers de Bruner, Goodnow et Austin (1956). Une catégorie est définie par une liste de propriétés nécessaires et suffisantes au sein de laquelle tous les objets sont considérés comme étant équivalents quant à leur appartenance catégorielle et le processus de catégorisation envisagé est un processus logique de découverte d'une règle de classification. L'univers des objets considérés résulte d'une combinatoire sur des valeurs de dimensions bien identifiées, indépendantes et manipulées dans un contexte expérimental. Très souvent les objets sur lesquels sont effectués les expériences sont des figures géométriques simples. L'expérimentateur choisit de manière arbitraire une règle de classification et pose l'expérience dans un cadre aux contours bien délimités et il s'agit alors de trouver les lois d'organisation en dehors de toute autre activité que celle de la logique. Un second cadre d'analyse, celui des catégories naturelles, impulsé par Rosch (1973, 1975) prend en compte

l'organisation des catégories et leur fonctionnalité. La thèse défendue est que les catégories ne sont pas des entités logiques limitées, auxquelles l'appartenance d'un item est simplement définie par le fait qu'il possède l'ensemble des propriétés répondants aux critères nécessaires et suffisants, et dont tous les cas qui possèdent ces critères sont également membres à part entière. La conception alternative proposée est que les catégories sont structurées par des effets prototypiques, déterminant des espaces catégoriels hétérogènes, caractérisés par des cas centraux typiques et des limites (Rosch, 1976). Ainsi, la catégorie se définit en référence à un prototype, soit le meilleur représentant de la catégorie. Le prototype condense l'ensemble des propriétés de la plupart des items, en fonction du principe d'économie cognitive. Cette condensation de la représentation de catégories sous formes de prototypes réduits les coûts de traitement, s'effectue de manière globale et permet des inférences sur des valeurs par défaut. Aussi, les prototypes correspondent aux exemplaires les plus fréquemment cités, les plus rapidement identifiés, les plus disponibles.

L'étude de la catégorisation permet d'établir quels sont les critères et les types de traitements utilisés par le système cognitif pour catégoriser et de rendre compte de la manière dont les connaissances sont structurées dans la mémoire sémantique à long terme.

2.1.5 Le processus "coarse to fine"

Des mesures provenant de travaux en biologie et en psychologie cognitive permettent d'affirmer que le processus de reconnaissance de scènes ou d'objets est très rapide chez l'homme (moins de 150 ms) (Torralba, 2003). Compte tenu du taux de transfert de l'information à travers les axones du cortex, et de contraintes biologiques sur les temps de latence entre les aires corticales, un schéma « en série » du traitement de l'information visuelle semble impossible. Le modèle suivant permet de modéliser la reconnaissance rapide de scènes ou d'objets :

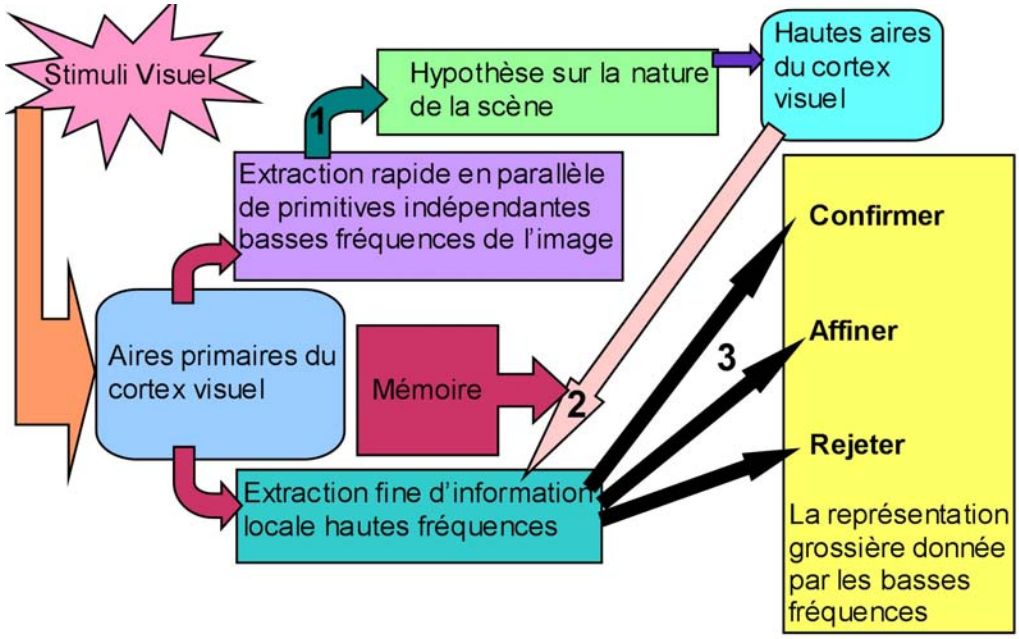

Figure 1 - Processus coarse-to-fine expliquant la reconnaissance rapide de scène chez l'homme

Lorsque la scène se projette sur les premières aires corticales, un modèle grossier de représentation de la scène est extrait rapidement. Ce modèle utilise alors essentiellement les composantes basses fréquences de l'image. Ces primitives sont indépendantes entre elles et sont envoyées aux cellules complexes du cortex inféro temporel qui codent la vision globale. Un signal de retour est alors envoyé au cortex primaire. Ce signal contient des hypothèses sur la nature de la scène visualisée, compte tenu des informations basses fréquences (BF) rapidement extraites et d'informations en provenance de la mémoire. Une seconde analyse de la scène est alors faite, visant à extraire de façon plus fine des informations locales provenant des hautes fréquences (HF). Cette information locale est ensuite utilisée pour affiner, confirmer ou rejeter la représentation grossière de la scène (basée sur les informations BF).

A partir de ce modèle, plusieurs théories sur la représentation et la reconnaissance automatique (donc par une machine) de scènes ou d'objets se sont développées ces dernières années. Dans un premier temps, c'est la reconnaissance par modèle qui a mobilisé les chercheurs. Plus récemment, les connaissances acquises sur le codage des scènes dans le cortex visuel des primates ont permis l'émergence de la reconnaissance par extraction de critères ou primitives. Nous ne nous étendrons pas plus sur les substrats neuronaux et cognitifs de la catégorisation chez l'homme. Nous allons maintenant nous intéresser aux différentes techniques de traitement et d'analyses des images.

2.2 Traitement du signal et analyses automatiques des images

Catégoriser des images consiste à les répartir automatiquement entre deux ou plusieurs catégories distinctes ou communes. Le classement automatique peut ensuite être comparé à un classement humain pour estimer ses performances.

2.2.1 Généralités et définitions

Le traitement du signal est la discipline qui développe et étudie les techniques de traitement (filtrage, amplification...), d'analyse et d'interprétation des signaux. Elle fait donc largement appel aux résultats de la théorie de l'information, des statistiques ainsi qu'à de nombreux autres domaines des mathématiques appliquées.

De nombreuses recherches se sont attachées à trouver des descripteurs dits de bas niveaux, comme les textures, les couleurs, les orientations locales ou les spectres afin de déterminer les catégories d'images, c'est-à-dire en induire les attributs de plus hauts niveaux (Guérin-Dugué & Oliva, 2000; Oliva, Torralba, Guérin-Dugué & Hérault, 1999 ; Vailaya, Jain & Zhang, 1998 ; Szummer & Picard, 1998).

Le traitement d'images désigne en informatique l'ensemble des traitements automatisés qui permettent, à partir d'images numérisées, de produire d'autres images numériques ou d'en extraire de l'information.

Dans le contexte de la vision artificielle, le traitement d'images se place après les étapes d'acquisition et de numérisation, assurant les transformations d'images et la partie de calcul permettant d'aller vers une interprétation des images traitées. Cette phase d'interprétation est d'ailleurs de plus en plus intégrée dans le traitement d'images, en faisant appel notamment à l'intelligence artificielle pour manipuler des connaissances, principalement sur les informations dont on dispose à propos de ce que représentent les images traitées, ce que l'on appelle connaissance du domaine (Schoenauer, 2001).

2.2.2 Recherches d'images et indexation de contenu

La recherche d'images par le contenu, en anglais : Content Based Image Retrieval (CBIR), est une technique visant à effectuer des recherches d'images à l'aide de requêtes portant sur les caractéristiques visuelles d'une image : texture, couleur, forme... à partir d'un moteur de recherche. Le cas typique d'utilisation de ces systèmes est lorsque l'on dispose d'une image pour laquelle on souhaiterait obtenir des images visuellement similaires. Il s'oppose à la recherche d'images par mots clés, qui est typiquement ce qui est proposé actuellement par les moteurs de recherche tels que Google ou Yahoo, où les images sont retrouvées en utilisant le texte qui les entoure plutôt que le contenu de l'image elle-même.

Les CBIR tentent, à l'inverse, de permettre une indexation et une recherche de l'image portant sur les caractéristiques de l'image : La texture (utilisation de filtres de Gabor, transformées en ondelettes discrètes, ...), la couleur (utilisation d'histogrammes dans l'espace RGB, TSV, ...), les formes (descripteurs de Fourier, ...), une combinaison de plusieurs de ces caractéristiques.
Ces caractéristiques sont dites de bas-niveau, car elles sont très proches du signal, et ne véhiculent pas de sémantique particulière sur l'image. Une fois ces caractéristiques extraites, la suite consiste généralement à définir diverses distances entre ces caractéristiques, et de définir une mesure de similarité globale entre deux images. Armés de cette mesure de similarité et d'une image requête, on peut alors calculer l'ensemble des mesures similarités entre cette image requête et l'ensemble des images de la base d'images. On peut alors ordonner les images de la base suivant leur score, et présenter le résultat à l'utilisateur, les images de plus grand score étant considérées comme les plus similaires.
Du fait des caractéristiques calculées, qui sont de bas-niveau, ces techniques obtiennent des résultats satisfaisants pour certains types de requêtes et certains types de base d'images. Par exemple rechercher des images de paysages enneigés, parmi une base d'image de paysages.
Toutefois ces systèmes rendent souvent des réponses extravagantes, et souvent éloignées de l'idée qu'avait l'utilisateur lorsqu'il a soumit sa requête.
Ce genre de système permet aussi de rechercher des images sans forcément avoir une image requête, par exemple rechercher des images plutôt bleues, ou alors dessiner une forme et demander de chercher toutes les images qui possèdent un objet de forme similaire.
Il existe plusieurs prototypes implémentant ce genre de techniques. Selon les critères de catégorisation utilisés, les systèmes sont capables de faire des différenciations diverses : scènes artificielles (villes, banlieues…) versus scènes naturelles (paysages de plaines, de montagnes, de forêts…), scènes d'intérieur (pièces de maison, de hangars, de bureaux) versus paysages extérieurs (villes, paysages naturels…), scènes ouvertes (c'est-à-dire présentant une ligne d'horizon) versus scènes fermées (sous bois, montagnes…), (Mojsilovic & Rogowitz, 2001). Certains systèmes récents différencient plus de deux catégories simultanément.
Le domaine fait toutefois encore partie de la recherche et n'est pas encore considéré comme mature.

2.2.3 Les différentes techniques d'analyses d'images

2.2.3.1 Hautes et basses fréquences spatiales

La plupart des techniques d'analyses d'images utilisent une décomposition en hautes et basses fréquences spatiales. La fréquence spatiale est l'inverse de la distance angulaire. Une image est un signal bidimensionnel que l'on peut décomposer selon ses fréquences spatiales.

Basses fréquences spatiales (BF): Les basses fréquences spatiales codent les formes dans leur globalité. Voici une illustration schématique de BF 

Hautes fréquences spatiales (HF): Les hautes fréquences spatiales codent les détails et les formes plus précises des objets dans la scène.

Voici une illustration schématique de HF 

L'utilisation des fréquences spatiales est possible pour des résolutions très faibles sans avoir d'informations sur l'identification précises des objets, elle est de plus relativement indépendante de la localisation des objets dans la scène.

L'ensemble des informations des basses et hautes fréquences spatiales peut définir une carte de saillance spatiale liée à l'analyse des zones d'intérêt de l'image.

Ainsi les expériences de psychophysique montrent qu'avec une présentation très courte, les sujets font peu d'erreur pour reconnaître les scènes naturelles (Marendaz, 1996; Torralba, 1999). Les résultats psychophysiques de catégorisation des scènes naturelles (Hérault et al, 1997; Oliva et al, 1999) montrent qu'il est possible, par calcul numérique, de catégoriser des scènes d'un point de vue sémantique, en se servant uniquement des basses fréquences spatiales. Les villes possèdent plus de hautes fréquences, en raison de plus de contrastes, les forêts plus de basses fréquences. Des résultats analogues existent en analysant les statistiques globales de la distribution des orientations dans une image (Vailaya et al, 1998; Guérin-Dugué, Oliva, 2000), dans ces paradigmes, la résolution optimale pour la catégorisation se situe plutôt vers les basses fréquences spatiales.

2.2.3.2 Autres techniques d'analyses automatiques d'images

Il existe quantités de techniques d'analyses d'images que je ne détaillerais pas ici. Citons l'analyse statistique globale contre locale, l'analyse chromatique et spatio-chromatique, l'analyse spectrale, les statistiques d'orientations des images naturelles, les filtres de Gabor, l'analyse en composante principal…

2.3 Paradigmes expérimentaux de classification d'images

De nombreuses études ont été consacrées à l'identification, à la reconnaissance ou à la classification d'objets simples, souvent dessinés au trait. En revanche celles concernant la perception de stimuli complexes tels que des images naturelles, paysages urbains ou champêtre, sont plus difficile à réaliser. Certaines études utilisent fréquemment des stimuli visuels représentant des scènes complexes stylisées, dessinées au trait (Palmer, 1975, Biederman, Mazzenote et Rabinowitz, 1982). L'utilisation de photos, plus proche de la réalité perceptive, concerne souvent la perception de visages, parfois celle de scènes visuelles complexes représentant des paysages urbains, plus rarement des paysages de campagne (Biederman, 1972; Potter, 1975; Biederman et Yu, 1988; Schyns et Oliva, 1994). Les arguments épistémologiques justifiant une démarche analytique dans l'étude de la perception visuelle furent sévèrement critiqués par Gibson (1979), à cause de ce que l'étude des situations "écologiques", défendue par Gibson, est souvent associée à l'observation phénoménologique et s'inscrit difficilement dans le cadre des théories du traitement de l'information.

De nombreuses études se sont intéressées à classer des images naturelles. Nous passons ici en revue les plus marquantes, ainsi que leurs méthodes.

2.3.1 Paradigme de Vailaya et al. (1998)

Vailaya et al. (1998) demandent à huit sujets humains de faire des catégories d'images, en étant libres des critères qu'ils utilisent, et en disposant du temps nécessaire. Les sujets séparent 171 images en douze catégories en moyenne. Les auteurs fabriquent ensuite une matrice de dissimilitude entre les images à partir de cette expérience et établissent un dendrogramme entre les images puis entre les catégories trouvées. Cela leur permet de définir une organisation hiérarchique des images contenues dans leur base. Les images qui se retrouvent immédiatement séparées sont les 'paysages', les 'villes' et les 'visages'. Les catégories 'paysages' et 'villes' sont elles-mêmes divisées en plusieurs autres catégories. Les auteurs essaient alors de reproduire certaines de ces catégories à partir de différents descripteurs liés à la couleur, aux fréquences ou aux directions de bords prépondérantes dans les images. En choisissant bien les classes et les descripteurs associés, ils atteignent des taux de classification de l'ordre de 94% pour la discrimination de deux classes.

2.3.2 L'expérience de Rogowitz (1998)

L'expérience de Rogowitz (1998) porte sur la mesure de similarité entre images, et se déroule en deux temps. Les images choisies couvrent un certain nombre de catégories

(paysages, villes, êtres vivants…). Les 97 images sont en couleur et sont prises avec des angles de vue différents.

Dans une première étape, appelée « *Table Scaling* », neuf sujets doivent organiser les images sur une table ronde, de telle sorte que celles qu'ils disposent éloignées sur la table soient des images qu'ils jugent très "différentes", et celles qu'ils disposent proches soient celles jugées "similaires". Les sujets projettent sur la table les 97 images. A la fin de cette étape, les auteurs disposent donc d'une matrice de distance entre leurs images (ils moyennent les distances obtenues entre toutes les images à la fin de chaque session).

Dans une deuxième étape, appelée « *Computer Scaling* », les sujets (quinze) jugent encore de la similitude entre les 97 images mais celles-ci sont visualisées sur l'écran d'un ordinateur. A chaque essai, une image à gauche de l'écran est présentée en face de huit autres images. Les sujets doivent choisir parmi ces huit autres images (prises au hasard parmi les 96 restantes) laquelle leur semble la plus proche de l'image de gauche. A la fin de l'expérience, les auteurs disposent donc d'une matrice de similarité entre les images ; cette matrice n'est pas complète car toutes les comparaisons entre images n'ont pas été présentées. Les auteurs demandent également aux sujets les critères qu'ils ont utilisés pour faire leurs associations.
Dans les deux expériences, les sujets sont libres de choisir le critère qui leur semble le plus approprié pour juger de la similarité entre les images.

Au final, les auteurs projettent ensuite les deux matrices, obtenues à la suite des deux expériences, en utilisant des algorithmes de MDS (Multidimensionnal Scaling). Ils comparent ensuite les deux projections obtenues à partir de ces "mesures perceptives", à la projection obtenue avec des matrices de distances entre les images, calculées à partir de descripteurs comme les histogrammes couleurs, ou un descripteur combinant couleur et orientation. Ils tentent par ces comparaisons de trouver les descripteurs mathématiques à partir desquels ils retrouvent une organisation entre images similaire à celle faite par les sujets, espérant ainsi trouver les descripteurs utilisés par les sujets dans des tâches de perception.
Deux dimensions dans l'organisation des scènes naturelles ressortent principalement l'axe "présence d'êtres vivants – pas d'êtres vivants" et l'axe "naturel – artificiel".
En 2001, Mojsilovic et Rogowitz déterminent un certain nombre de catégories qu'elles appellent sémantiques parmi les images de scènes naturelles. Ainsi les catégories suivantes sont dégagées : les portraits, les personnages dans des scènes d'intérieurs, les personnages dans des scènes d'extérieurs, les scènes d'extérieurs avec des personnages, la foule, les villes, les bâtiments, les « technoscènes » (des péages d'autoroutes, des pylônes électriques, etc), des

objets à l'intérieur, des objets à l'extérieur, le ciel, la neige, les animaux, les textures, les paysages de verdure, les paysages avec de l'eau.

2.3.3 Paradigme de l'ordonnancement

Ce paradigme proposé par Oliva, Torralba, Guerrin-Dugué, Hérault, en 1999. Son but est l'organisation par les sujets d'une base de 470 images en 4 catégories (Outdoor, Indoor, Closed, Open). On remarque un certain regroupement des images appartenant à la même catégorie. On peut également extraire les axes le long desquels les images vont s'ordonner. Axes facilement interprétables par des sujets humains. Il est important de noter que Oliva et al. (1999) parlent de la nécessité de définir des axes avec une organisation continue des images et non pas de rester avec des classes possédant des frontières strictes. En effet, sur un grand nombre d'images différentes, il se peut qu'une image soit difficilement classable dans une seule catégorie ; par exemple, une image de montagne au bord de la mer se situe entre la classe des « montagnes » et celle des « plages ». C'est pourquoi il semble pertinent, de trouver des descripteurs qui offrent une organisation de nos images en continuum, par rapport à une organisation uniquement en « cluster ».

2.3.4 Le paradigme de Guérin-Dugué et Oliva (2000)

Guérin-Dugué et Oliva (2000) utilisent ici encore les orientations des spectres des images afin de classer les images cette fois-ci en quatre catégories, qu'elles appellent sémantiques : les scènes d'intérieurs, les scènes urbaines, les paysages et les scènes ouvertes (Guérin-Dugué & Oliva, 2000). Elles utilisent les caractéristiques des spectres de 470 images en niveaux de gris. A chaque point de l'image, sont associées son orientation dominante et son énergie à travers des filtres orientés et cela pour différentes fréquences spatiales. Ces informations sont regroupées sous forme d'histogramme. Une image est donc décrite par un certain nombre d'histogrammes, un pour chaque fréquence spatiale. Puis, par la méthode des plus proches voisins (distance euclidienne), les images sont catégorisées suivant quatre classes. Les auteurs obtiennent des taux de catégorisation compris entre 82% et 90%.

2.4 Conclusion sur la bibliographie

Comme nous venons de le voir, les modèles et analyses, qui s'inspirent de la perception visuelle, dans le but de catégoriser et de faire des systèmes de recherches d'images par le contenu, sont riches et variés, et visent tous à faciliter l'interaction entre le système et l'utilisateur dans sa démarche de recherche d'image. La similarité entre les images vues et recherchées par l'utilisateur en quête d'information n'est pas accessible, car on ne connaît pas sa vraie perception, sa vraie motivation et son interprétation de l'image, elle est subjective.

La catégorisation de scènes offre un cadre expérimental bien défini et souvent exploité permettant de définir les principales catégories. Mais en pratique, ces paradigmes paraissent peu exploitables pour isoler les critères de catégorisation, en raison de situations très diverses, dans de grandes bases d'images, où beaucoup d'éléments varient, y compris à l'intérieur d'une même catégorie. Les images proposées au sein des différents paradigmes expérimentaux couvrent une large gamme d'environnements possibles. Introduire dans un paradigme une mesure précise des critères de catégorisation pour l'extraction d'attributs d'image permettra d'améliorer la fiabilité des systèmes de recherche d'information, et le degré de satisfaction de l'utilisateur. Il faut donc, pour répondre aux questions posées se concentrer uniquement sur la recherche de descripteurs permettant de catégoriser selon des critères humains les images. Ensuite il faudra utiliser les différentes techniques de traitement du signal, pour paramétrer le moteur de recherche, afin que son comportement soit le plus proche possible du sujet humain moyen.

3 Protocole

Comme nous l'avons vu précédent, les protocoles déjà existants ne paraissaient pas satisfaisants pour répondre efficacement aux questions posées. Il fallait donc créer un nouveau protocole.

L'esprit du protocole réside dans l'intention de faire varier des éléments de chaque image en préservant le principe de "toutes choses étant égales par ailleurs", néanmoins, la nature même des stimuli, en tant qu'image complexe, complique la tâche. Nous avons décidé pour réduire quelque peu la complexité de images, de n'y mettre qu'un contenu sémantique limité – seulement des villes et des forêt- en préservant leur aspect naturel. Pour cela les images ont dû être modifiées à l'aide de Photoshop CS2, suffisamment pour supprimer les variables parasites ostentatoires, mais assez peu pour qu'elles gardent leur aspect naturel, celle de vraies photos non retouchées.

3.1 Variations des éléments – principes généraux

Ainsi pour chaque image nous avons choisi de faire varier certaines variables, selon des critères, suspectés comme pertinents dans l'activité de catégorisation humaine. Ainsi chaque image est composée d'un élément dominant, ici toujours ville ou forêt. L'élément dominant est toujours l'élément sur lequel s'exerce la variation. Ainsi la première lettre de chaque image représente l'élément en question, ici ville (V) ou forêt (F). La figure 2 montre le principe de l'élément dominant.

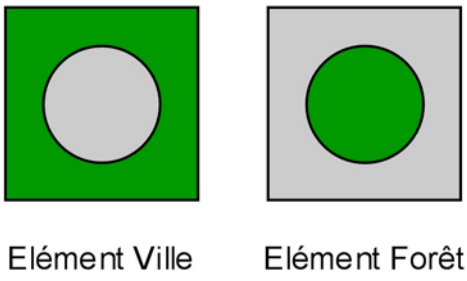

Figure 2 - Principe de l'élément dominant

Au niveau perceptif, les villes présentes à l'image laissent voir une organisation géométrique des objets les constituants, des lignes nettes délimitant des formes non accidentelles. Elles sont plus contrastée, et de ce fait possèdent plus de hautes fréquences. Les forêts, quand à elles, montrent des contours ne représentant pas une organisation géométrique, mais des formes accidentelles, dans laquelle une organisation basse fréquence est plus présente.

3.2 Les Variables Indépendantes (VI)

3.2.1 – VI 1 Le centrage

La première variable concerne le centrage ou le décentrage de l'élément dominant (toujours ville ou forêt dans le protocole). Elle comporte deux modalités "Centré – C" ou "Décentré – D". Ainsi un élément centré se situera approximativement au centre de l'image, sans avoir de frontière commune avec le bord de la photo. Un élément décentré sera toujours en contact avec un des bords de l'image, et n'occupera la zone centrale de l'image que si sa proportion le justifie. La seconde lettre de chaque image est toujours soit un C, pour élément Centré, soit un D, pour élément Décentré.

3.2.2 - VI 2 La proportion

La seconde variable concerne la proportion de l'élément dominant dans l'image. Le nom de chaque image est constitué d'un chiffre, en troisième position. Ce chiffre représente la proportion de l'élément dominant dans l'image. Il représente un continuum entre 01 et 90, et comporte trois modalités "Petit" pour les chiffres entre 01 et 25, "Moyen", pour les chiffres entre 26 et 49, et "Grand" pour les chiffres de 50 à 90. Ces chiffres représentent la proportion présumée de la taille de l'élément dans l'image.

Les figures 7 et 8 montrent schématiquement la variation des images en fonction de la VI 1 et de la VI 2. On peut ainsi voir l'élément dominant, ville ou forêt, changer de position et de proportion. Pour des illustrations plus concrètes voir l'Annexe.

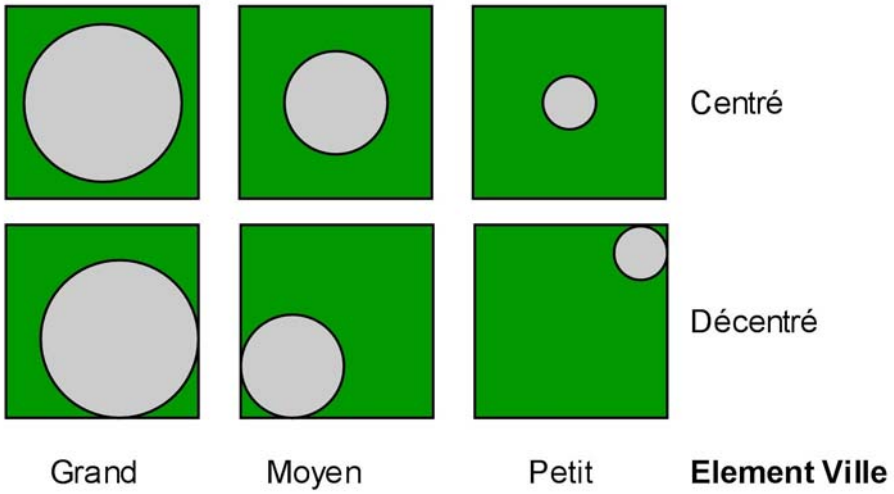

Figure 3 - Schéma des images créées. Elément ville

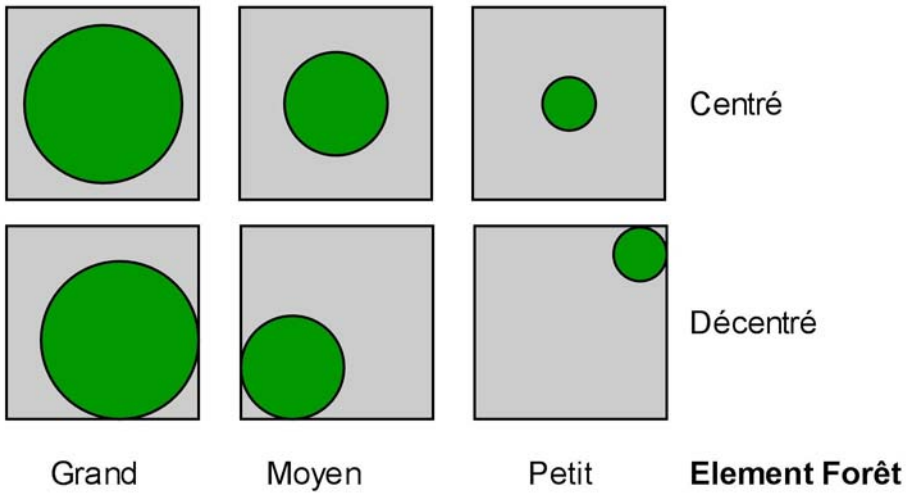

Figure 4 - Schéma des images créées. Elément forêt

3.2.3 – VI 3 Le filtrage

Le filtrage est la troisième variation, qui s'applique sur l'ensemble de l'image. Cette variable est créée afin de savoir si l'usage des filtres modifie l'activité de catégorisation ou si ils n'ont pas un effet significatif. La variable "filtrage" comporte trois modalités. Haut "H", Bas "B" et Neutre "N" (aucun filtrage), modalité contrôle de cette variable. La dernière lettre de chaque image est corrélativement la lettre H, B, ou N.

3.3 Codage des images

Chaque image possède un nom. On peut retrouver les caractéristiques des images, les VI dans leur nom. Ainsi l'image FD30H, est une image dont l'élément dominant est une forêt, qui est

décentrée, dont la proportion est approximativement de 30%, avec un filtrage passe haut. Ainsi on peut très facilement retrouver les caractéristiques des images.

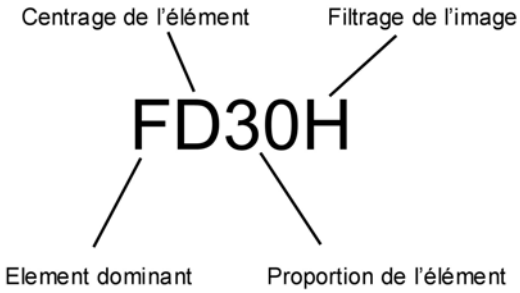

Figure 5 - Principe de codage des images

3.4 Groupes et sujets

Avant d'aller plus loin dans l'explication du protocole, il est important que nous abordions la thématique des groupes dans l'expérience. Ici il n'y a pas besoin de groupes témoins, juste d'échantillons les plus vastes et représentatifs possibles. Le classifieur lorsqu'il est face à une image ambiguë, peut prendre la décision de classer l'image dans plusieurs catégories. Ainsi l'image sera "forêt ET ville". Ce choix peut s'avérer pertinent, mais pas toujours. En effet un arbre isolé dans une ville, ne devrait pas être classifié comme "forêt", à l'instar d'une maison dans les bois ne pourrait être classifiée en "ville". La question que l'on se pose ici est de savoir à partir de quand une ville devient une ville. Il faut donc que les sujets aient l'opportunité de répondre "les deux". Afin qu'ils puissent réagir comme la machine, les sujets ont trois choix, "ville", "forêt", "les deux". C'est un choix obéissant à la fonction logique OR inclusif. Il forme le premier groupe, appelé groupe OR. Nous avons choisi de créer un second groupe, parce que la décision de choix catégoriel à trois possibilités est peu explorée en sciences cognitives, et qu'elle est sans doute porteuse de biais. Ainsi la présence d'un second groupe ne possédant qu'un choix obéissant à la fonction logique XOR, c'est-à-dire à deux choix, est sans conteste mieux maîtrisée, et peut permettre d'apporter des éléments de comparaison intéressants sur les différentes variables dépendantes, ainsi que sur les images catégorisées comme appartenant à la catégorie "les deux" par les groupe OR. Ce deuxième groupe sera appelé groupe XOR.
Dans chaque groupe il y a 20 sujets, soit un total de 40 sujets.

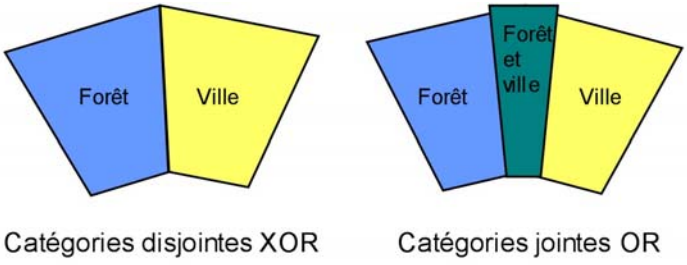

Figure 6 - Logique des groupes XOR et OR

3.5 Plan expérimental

Le plan expérimental est du type $\mathbf{S_{20} <G_2<C_2*P_3*F_3>>}$

Où S est le nombre de sujet, G le nombre de groupe, C le centrage, P la proportion, F le filtrage. C est une variable nominale, P est une variable ordinale, et F est une variable nominale.

3.6 Les Variables Dépendantes (VD)

Les variables dépendantes représentent ce qui est mesuré dans l'expérience. Il en existe trois dans le protocole.

3.6.1 Choix catégoriel

La première VD est le choix catégoriel des sujets. C'est la mesure la plus importante. Elle est nominale. Elle est constitué de deux modalités ("ville" ou "forêt") dans le groupe XOR, et de trois modalités ("ville", "forêt", "les deux") dans le groupe OR. C'est la VD principale.

3.6.2 Temps de Réponse (TR)

Le temps de réponse est la seconde VD. La mesure du TR est une méthode classique des sciences cognitives. Elle est corrélée aux différentes étapes de traitements mis en œuvre par les sujets pour réaliser leur tache. Le TR ne s'applique ici qu'à la décision de jugement catégoriel, et non au jugement de typicalité.

3.6.3 Typicalité

La seconde VD est la typicalité. Elle est secondaire, et n'a qu'une valeur de possibilité de rectification en cas d'erreur dans leur classification. Si ils répondent 1, l'appartenance de l'image à la catégorie est rejetée. Ainsi il est demandé de définir typiquement les images, pour rythmer le protocole. Si les sujets choisissent les catégories strictes "forêt" ou "ville", l'ordinateur leur demandera si l'image est peu ou très représentative. Dans le groupe OR, si les sujets choisissent la catégorie "les deux" l'ordinateur leur demandera si l'image est tout de même plus une "forêt" qu'une "ville" ou l'inverse. La réponse typicalité n'est pas chronométrée. C'est une variable ordinale, composée de cinq positions de "Très peu typique" à "Très typique".

3.7 Variables confondues

En réalité les villes ne sont pas toutes rondes, bien centrées, au milieu d'une forêt d'une densité égale et bien homogène. Ces schémas sont des images simples, et les photos que le classifieur doit catégoriser sont des images complexes, de tailles, et de définitions très hétérogènes. Les villes ont des allures non semblables, composées d'immeubles, de maisons,

de rues, d'aspects très différents. Les forêts ont un aspect plus homogène, même si leur aspect peut également beaucoup varier, en raison des espèces, de la densité, de la saison… Il existe donc intrinsèquement à chaque image, de par sa complexité, quantités de variables dites confondues, qui ne sont pas ici prise en compte, car trop nombreuses, non conceptualisées, ni même imaginées. Créer des images en conservant le principe de "toutes choses étant égales par ailleurs" n'est ici pas totalement possible. Un des biais de l'expérience, est qu'en voulant préserver ce principe, les images se ressemblent beaucoup, et les sujets peuvent faire appel à leur mémoire pour répondre. Nous n'avons à ce jour pas trouvé de méthode adéquate pour contrebalancer ce biais, tout en conservant le principe initial. Chaque image a été modifiée grâce au tampon de Photoshop, pour unifier autant que faire ce peu la densité des éléments, mettre des nuances de verts identiques pour les forêts, supprimer des parties de ville présente à l'extérieur des centres urbains, transformer les champs environnants en forêts …

Il existe ainsi pour chaque image, de par sa nature complexe quantités de variables confondues, c'est-à-dire des éléments qui rentrent en compte dans l'activité de catégorisation, mais qui ne sont pas identifiés.

3.8 Le programme

Le logiciel du protocole est réalisé avec Visual C#.net. Il est présent sur le CD-ROM en annexe. Il a été conçu afin que les images défilent selon un mode aléatoire. Les touches du clavier permettant la réponse sont contrebalancées par le programme, qui modifie l'attribution des boutons à chaque nouveau sujet, afin d'éviter un biais. Une phase d'entraînement a été prévue afin que les sujets se familiarisent avec l'interface. Tous les éléments du protocole ont été incorporés au programme.

3.9 Les sujets

Nous avons choisis de faire passer 20 sujets par groupe, soit 40 au total. Tous les sujets ont entre 18 et 35 ans, il y a autant d'hommes que de femmes dans chaque groupe. Chaque sujet a une bonne vue, sans ou avec correction. Le programme leur est présenté sur une table avec un ordinateur portable DELL XPS M1210.

3.10 Les hypothèses

Nous avons formulé cinq hypothèses intuitives :

1 * La proportions interagit avec l'activité de catégorisation.

2 * Le centrage interagit avec l'activité de catégorisation.

3 * Le filtrage n'interagit pas avec l'activité de catégorisation.

4 * La proportions interagit avec le TR.

5 * Le centrage interagit avec le TR.

Ainsi le protocole sera chargé de mettre en évidence ces effets là, et de voir si les hypothèses sont confirmées ou infirmées.

4 Résultats et analyses

Tous les résultats ont fait l'objet d'une ANOVA, qui a permis de détecter les effets principaux, et les interactions entre facteurs. La variance dans les différents groupes soit la même

4.1 Analyses de l'activité de catégorisation

Nous allons analyser les effets des variables indépendantes sur la variable dépendante catégorisation.

4.1.1 Analyses des effets principaux

4.1.1.1 Proportion

L'effet principal Proportion apparaît significatif après analyse de la variance, $F(2,38)= 32.4$, $p<.001$, pour le groupe XOR, et $F(2,38)= 15.48$, $p<.001$ pour le groupe OR.

Ces résultats sont illustrés par la figure 7, qui nous montre l'effet significatif de la proportion sur la catégorisation, dans un contexte de choix exclusif.

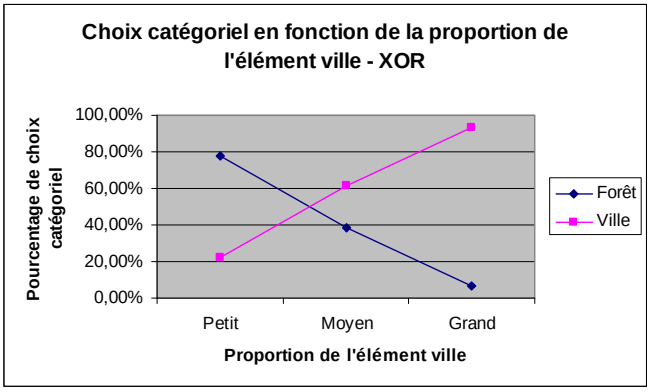

Figure 7 – Choix catégoriel en fonction de la proportion de l'élément ville - XOR

On peut donc constater que lorsque le choix catégoriel des sujets varie en fonction de la proportion de l'élément dans l'image. Ainsi plus l'élément a une taille importante, plus le sujet à tendance à choisir la catégorie s'y rapportant.

La figure 8, nous montre le même effet significatif, mais cette fois dans un contexte de choix inclusif. Ici, la possibilité de pouvoir répondre "les deux" est majoritaire, et trouve sont optimum lorsque l'élément ville est moyen, c'est-à-dire lorsque le choix est le plus ambigu.

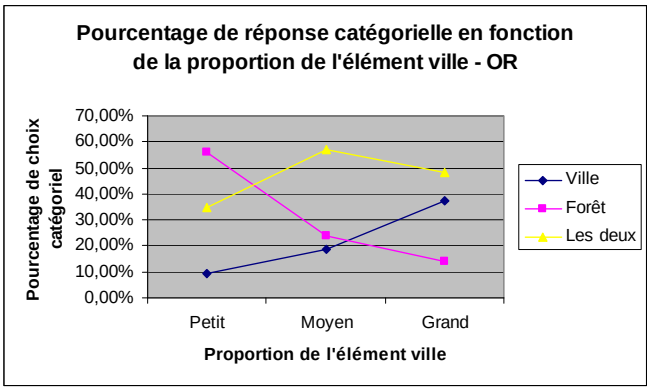

Figure 8 – Pourcentage de réponse catégorielle en fonction de la proportion de l'élément ville - OR

Sur la figure 9, la réponse "les deux" c'est-à-dire l'hésitation, trouve sont optimum avec une proportion moyenne, là où l'image est plus ambigu, et est plus basse pour des tailles petites et grandes, lorsque l'ambiguïté est moins forte.

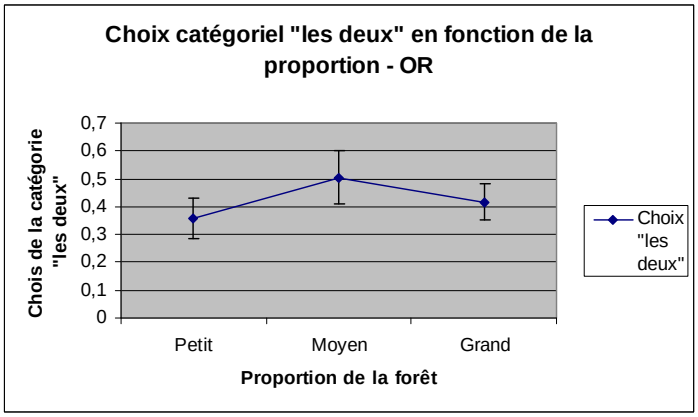

Figure 9 – Choix catégoriel, les deux, en fonction de la proportion - OR

4.1.1.2 Centrage

Nous allons maintenant étudier les effets principaux de la variable Centrage sur l'activité de catégorisation. L'effet principal Centrage apparaît significatif $F(1,19)= 29.9$, $p<.001$ pour le groupe XOR. La variable centrage n'est significative que pour les sujets catégorisant dans un contexte de choix exclusif. Le centrage semble permettre au sujet de trancher lorsque l'image est ambiguë, pour le groupe OR, l'effet du centrage n'est pas significatif, sans doute à cause du troisième choix "les deux", que les sujets choisissent si l'image est ambiguë.

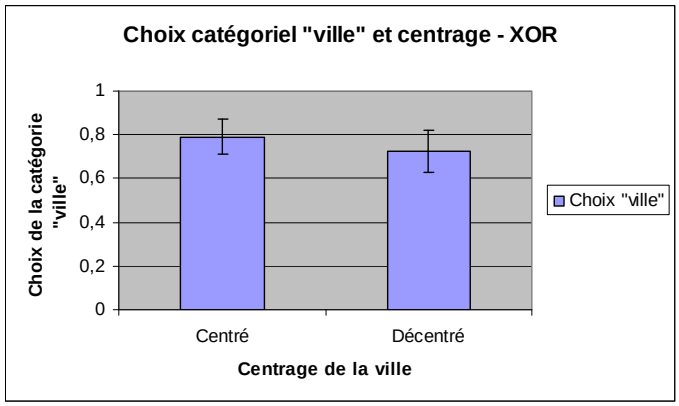

Figure 10 – Choix catégoriel, ville, et centrage - XOR

Sur la figure 10 on peut constater que lorsque l'élément ville est centré, nous avons un choix catégoriel plus prononcé en faveur de la ville. On observe le même effet lorsque c'est le centrage de la forêt qui varie, quand elle est décentrée, les sujets choisissent plus ville, et quand elle est centrée, ils choisissent plus forêt – figure 11.

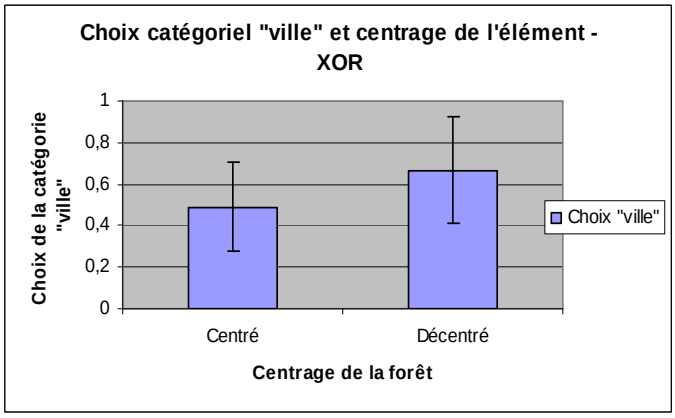

Figure 11 – Choix catégoriel, ville, et centrage de l' élément - XOR

4.1.1.3 Filtrage

Voici l'effet principal du Filtrage sur la catégorisation. Il est significatif sur le groupe XOR, $F(2,38)= 10.97$, $p<.001$. Par contre il ne l'est pas sur le groupe OR, qui possède trois choix. Nous savons (Cf. § 2.2.3.1), que les villes ont plus de hautes fréquences que de basses fréquences, en raison des forts contrastes, et qu'à l'inverse les forêts présentent plus de basses fréquences que de hautes fréquences. On peut observer, sur la figure 12, que les sujets choisissent plus la catégorie "ville" lorsque l'image a subi un filtrage haut. Ce qui montre que le filtrage joue un rôle dans l'activité de catégorisation. En ne laissant que les hautes fréquences, le filtrage augmente les contrastes des villes, et les rend plus saillante, ce qui a une influence sur le comportement des sujets.

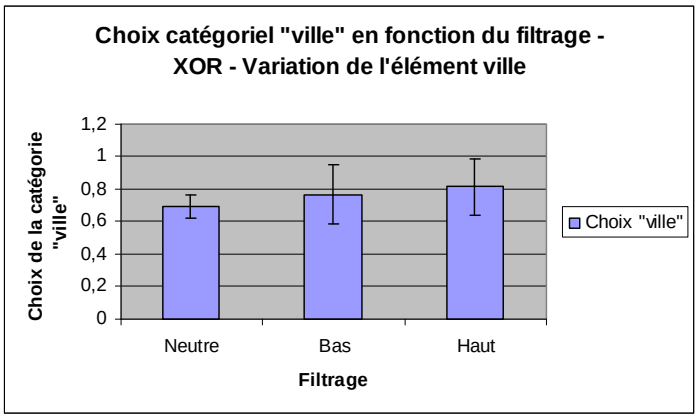

Figure 12 – Choix catégoriel, ville, en fonction du filtrage – XOR – Variation de l'élément ville

Les filtrages haut et bas semblent renforcer respectivement la saillance de la ville et de la forêt.

4.1.2 Analyses des effets croisés

Voici l'analyse des effets croisés des variables sur l'activité de catégorisation.

4.1.2.1 Filtrage * Proportion

L'interaction Filtrage * Proportion apparaît significative à l'analyse de la variance $F(4,76)= 20.64$, $p<.001$ pour le groupe XOR, et $F(4,76)= 12.84$, $p<.001$ pour le groupe OR.

On peut constater sur les figures suivantes, que les courbes filtrage haut et bas sont assez similaires. Par contre on constate que les courbes filtrage neutre sont différentes. Le filtrage neutre représente la condition contrôle.

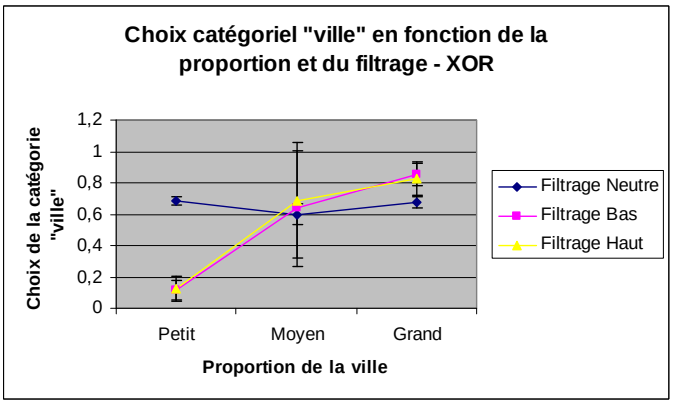

Figure 13 – Choix catégoriel, ville, en fonction de la proportion et du filtrage - XOR

Sur la figure 13, il est représenté le choix catégoriel "ville". Ce choix catégoriel semble renforcé, pour le filtrage haut et bas. Dans la condition contrôle le choix catégoriel est assez stable, et semble moins dépendre de la proportion de l'élément. On observe que le comportement de catégorisation des sujets change en fonction du filtrage, ainsi les sujets

catégorisent les images non filtrées presque de la même manière quelle que soit la variation de la taille de l'élément.

La figure 14 montre le choix catégoriel "les deux" des sujets, c'est-à-dire l'indécision. On peut observer que la courbe contrôle "filtrage neutre" est stable, et semble ne pas dépendre de la proportion de l'élément. Elle est en décalage avec les deux autres courbes qui sont assez identiques, et qui semblent être en corrélation étroite avec la proportion de l'élément.

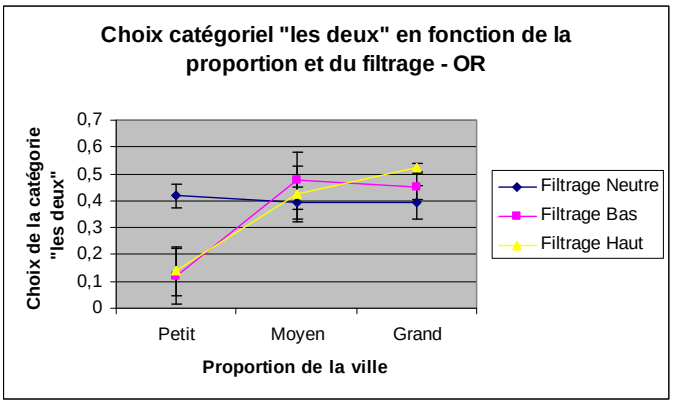

Figure 14 – Choix catégoriel, les deux, en fonction de la proportion et du filtrage - XOR

Nous pouvons donc dire que le filtrage augmente l'importance de la proportion, en renforçant la saillance des villes et des forêts. Il est néanmoins assez troublant que le filtrage neutre, ici condition contrôle, se démarque autant, et ait un rapport aussi stable avec la proportion.

4.1.2.2 Centrage * proportion

L'interaction Centrage * Proportion apparaît significatif à l'analyse de la variance $F(2,38)= 6.79$, $p<.01$.

On peut voir sur la figure 15 que le Centrage renforce fortement le choix catégoriel des sujets, aux alentours de 20% pour les proportions petite et moyenne. Avec une proportion grande, l'effet du centrage se fait moins sentir (environ 5%), parce que la grande taille de la ville réduit l'impact du Centrage, même décentrée la ville occupe le centre.

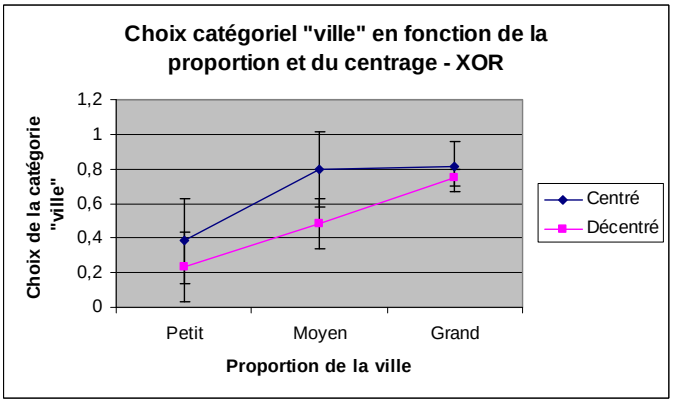

Figure 15 – Choix catégoriel, ville, en fonction de la proportion et du centrage - XOR

4.1.2.3 Centrage * filtrage

L'interaction Centrage * Filtrage apparaît significatif à l'analyse de la variance. $F(2,38)= 11.65$, $p<.001$.

On peut constater sur la figure 16, qu'un filtrage haut et bas renforce l'importance du centrage dans le choix catégoriel des sujets. Par contre le filtrage neutre, condition contrôle, réduit l'importance du centrage et incite les sujets à moins choisir la ville lorsqu'elle est centrée.

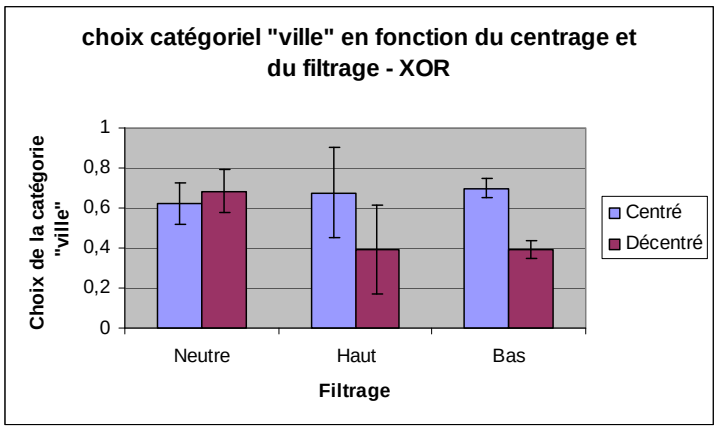

Figure 16 – Choix catégoriel, ville, en fonction du centrage et du filtrage - XOR

Ainsi le filtrage semble renforcer l'importance du Centrage dans l'image.

4.1.2.4 Centrage * proportion * filtrage

L'interaction triple Centrage * Proportion * Filtrage apparaît significative à l'analyse de la variance. $F(4,76)= 4.73$, $p<.01$. Nous pouvons constater la présence d'une triple interaction des trois facteurs sur l'activité de catégorisation. Pour représenter ce triple effet, nous utilisons deux graphiques en 2D – figure 17.

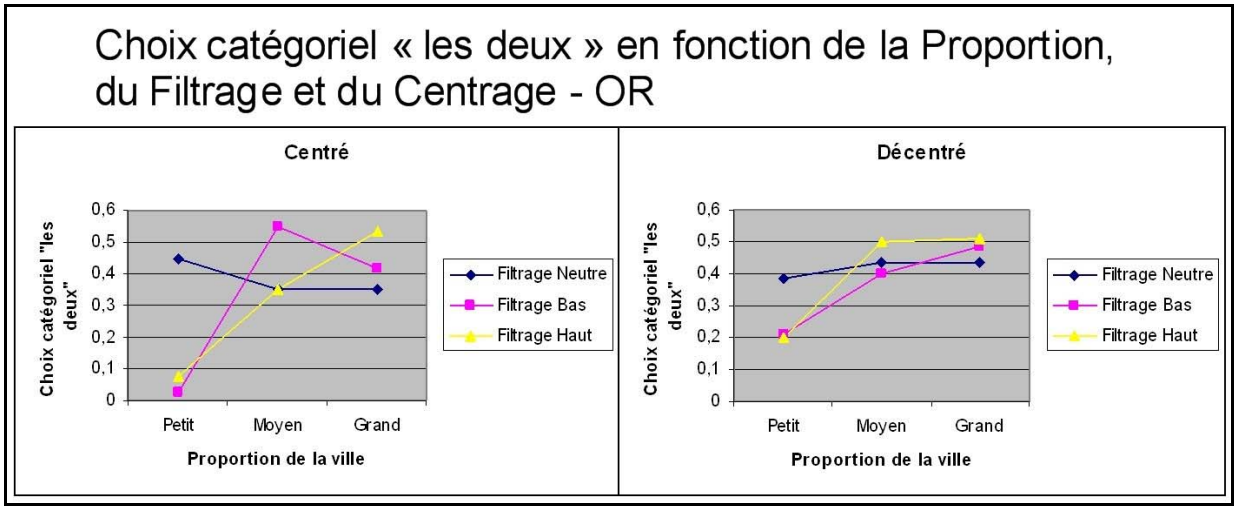

Figure 17 – Choix catégoriel, les deux, en fonction de la proportion du filtrage et du centrage - OR

Nous pouvons observer que lorsque l'élément ville est centré, l'effet du filtrage bas et haut augmente la saillance, et diminue l'hésitation à se prononcer sur la catégorie "les deux".

Lorsque l'élément ville est de taille moyenne, le filtrage bas augmente l'hésitation. Par contre lorsque l'élément ville est de grande taille, les filtrages hauts et bas augmentent l'hésitation, bien que ce soit à nuancer. Lorsque l'élément ville est décentré, et qu'il est de petite taille, les filtrages hauts et basses fréquences diminuent l'hésitation. Lorsque l'élément ville est de moyenne et grande taille, les filtrages n'ont pas vraiment un effet significatif. La courbe "Filtrage Neutre" se distingue des deux autres dans les deux modalités "Centrée" et "Décentrée", son comportement semble moins varier en fonction de la proportion. La courbe reste relativement stable. A l'inverse les courbes "Filtrage Haut" et "Filtrage Bas" ont un tracé qui semble plus sensible à la variable proportion. Les sujets humains répondent moins "les deux" lorsque la proportion est petite, et plus lorsque la proportion est "Moyen" ou "Grand". Le filtrage renforce cet effet, en mettant mieux en saillance les villes et les forêts. Lorsque l'image n'est pas filtrée, le taux de réponse "les deux", donc l'hésitation est assez stable. Le centrage de la ville augmente l'effet et la pente des courbes. Lorsque la ville est décentrée la pente des courbes est moins prononcée.

4.2 Analyses des TR

Maintenant nous allons analyser les effets des VI sur la variables dépendante temps de réponse, afin de savoir si il existe des interactions significatives ou des tendances.

4.2.1 Analyses des effets principaux

4.2.1.1 Proportion

L'effet principal Proportion apparaît significatif à l'analyse de la variance. $F(2,38)= 29.23$, $p<.001$. La figure 18 illustre cet effet. On peut y voir une variation du temps de réponse en fonction de la proportion de l'élément. De la proportion "petit" à "moyen", le temps de réponse augmente de presque une seconde, puis décroît légèrement à "grand". On peut également noter que les écarts types sont très resserrés, donc les réponses sont très homogènes.

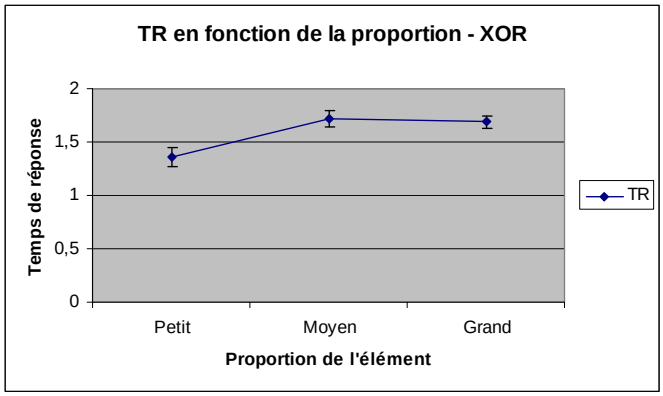

Figure 18 – TR en fonction de la proportion - XOR

On peut constater que le TR varie avec la proportion, sans doute en raison de l'ambiguïté des images, pour les écarts de proportion de "petit" à "moyen", ainsi qu'à cause de l'exploration visuelle, pour la proportion"grand". Une image comportant un élément de petite taille sera moins ambiguë qu'une comportant un élément de taille moyenne, et nécessitera moins d'exploration visuelle qu'une image avec un élément de grande taille. Ce qui explique l'allure générale de la courbe. Il existe une différence de temps de réponse entre les deux groupes. Dans le groupe OR, le temps de réponse est en moyenne plus long (TR moyen du groupe OR = 1,8 secondes (écart type moyen 0,78) contre 1,493 secondes pour le groupe XOR (écart type moyen 0,62), soit un écart de 0,31 secondes entre les deux groupes.)

4.2.1.2 Centrage

La variable Centrage seule ne semble pas avoir d'effets significatifs sur le TR.

4.2.1.3 Filtrage

La variable filtrage ne semble pas avoir d'effets significatifs sur le TR.

4.2.2 Analyses des effets croisés

4.2.2.1 Centrage * proportion

L'interaction Centrage * Proportion apparaît non significative à l'analyse de la variance. $F(2,38)= 3.07$, $p=.057$, mais il existe une tendance très proche du seuil de significativité.

On peut voir dans la figure 19, le TR est le plus bas lorsque la proportion de l'élément est "petit" et qu'il est décentré. Au niveau du TR cela se traduit par une facilité accrue de prise de décision. Par contre pour un élément de taille moyenne ou haute, décentré, le TR augmente, en raison de l'ambiguïté des images. Le sujet met donc plus de temps pour répondre. Il n'y a pas de différence significative entre une proportion moyenne et grande. Le TR des éléments

centrés est assez régulier. Nous pouvons expliquer ces résultats, en invoquant l'idée d'ambiguïté. Les TR des éléments décentrés de taille moyenne et grande sont légèrement plus longs car ils portent plus d'ambiguïtés, sauf pour une les éléments dont la proportion est "petit", car l'ambiguïté devient alors assez mince, et le choix plus facile.

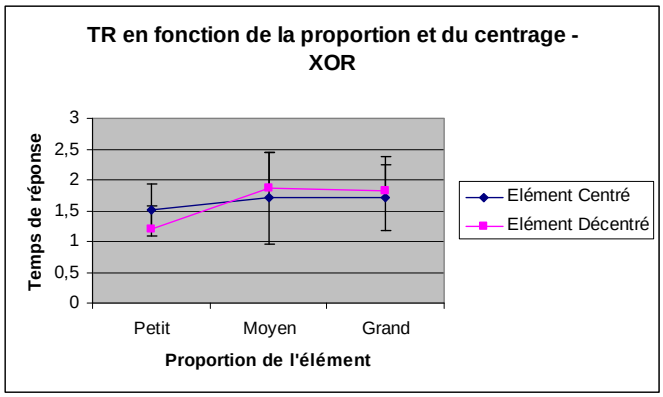

Figure 19 – TR en fonction de la proportion et du centrage - XOR

4.3 Conclusion des résultats et analyses

Nous avons eut dans cette expérience un certains nombres de résultats. Dans l'activité de catégorisation, la proportion joue un rôle important. Le centrage peut réduire l'ambiguïté des images. Le filtrage permet d'augmenter la saillance des éléments de l'image, l'importance de la proportion, et du centrage. A ce titre le filtrage a un rôle déterminant. Ces trois variables ont des effets significatifs sur la catégorisation. En ce qui concerne les temps de réponse, la proportion semble avoir un impact fort. Le TR est corrélé avec la proportion, ainsi qu'avec la proportion et le centrage, en rapport donc avec l'ambiguïté des images. Le filtrage, ainsi que le centrage n'ont pas avoir d'interactions significatives avec le TR. Les hypothèses que nous avons formulées sont toutes vérifiées à l'exception de la 3 et de la 5. La proportions interagi avec l'activité de catégorisation (Hypothèse 1 confirmée), le centrage interagi avec l'activité de catégorisation (Hypothèse 2 confirmée), le filtrage interagi avec l'activité de catégorisation (Hypothèse 3 infirmée), la proportion interagi avec le TR (Hypothèse 4 confirmée), le centrage interagi avec le TR (Hypothèse 5 confirmée).

5 Discussion

5.1 Sémantique non binaire des éléments de l'image

Une observation importante que l'on peut faire est propre à la situation expérimentale. En effet afin de simplifier au maximum l'activité de catégorisation, deux catégories seulement sont disponibles pour les sujets. Hors il s'avère que deux catégories sémantiques sont largement insuffisantes pour catégoriser ces images complexes. Ainsi les catégories "ville",

"village","hameau", "station", "cité", "bourgade" serait tout a fait pertinentes. Cette observation est sans doute encore plus important avec les catégories sous ordonnées de "forêt", comme "bois", "parc" (une forêt dans une ville), "bosquet", "boqueteau", "taillis" … Le robot classifieur se doit de catégoriser selon de nombreuses catégories, et la présence ici de seulement deux catégories inclusives ou exclusives n'est pas suffisante. Ainsi on peut constater que chez les sujets humains l'analyse sémantique est forte. La catégorie "ville" n'est pas toujours appropriée, mais plutôt "village" ou "station de montagne". La présence de seulement deux catégories est trop limitée face à la complexité perceptuelle et sémantique des images. De plus il semble exister un certains nombres de traitements propre à chaque catégories d'images, ainsi une image sera identifier comme "paysage de montagne" à la suite de traitement très différent de ceux pour identifier une "ville". Les critères isolés ici ne sont donc pas généralisables à l'ensemble des images.

5.1.1 Généralisation des résultats

Dans quelle mesure les résultats sont ils généralisables à l'ensemble des images ? Il apparaît que les catégories ne sont pas "vide de sens". Les catégories sont particulières, et que les traitements qui y sont attaché dépendent de ces catégories. Il serait donc probable, que les résultats seraient différents si nous avions choisis des catégories différentes au début de l'expérience, comme "forêt" et "montagne", ou "forêt" et "plage", ici que des catégories naturelles. Les résultats sont donc difficilement généralisables à l'ensemble des catégories. Il doit, de plus, exister des catégories super ordonnées, sur lesquelles s'appliquent des traitements particuliers. Il serait intéressant de prolonger cette étude afin d'identifier plus précisément ces catégories super ordonnées. Les catégories peuvent être classées en réseau, voir figure 25. A chaque niveau du réseau, les traitements sont différents, et dépendent du type d'image et de catégorie. On ne traite pas de la même manière une image d'une brouette ou de ciseaux, ou une image de plage, une image d'objet ou de symboles. Afin que le système puisse faire son travail sans trop utiliser de ressources, il convient de réaliser des traitements en cascade, qui s'affinent au fur et à mesure, jusqu'à arriver à une classification stable et précise de chaque image. Avec ce principe nous pouvons rêver à un classifieur universel.

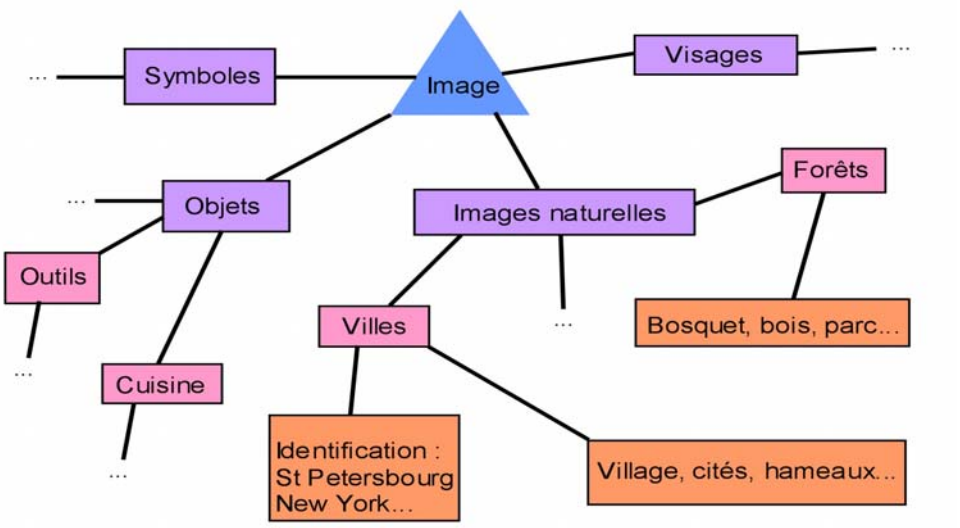

Figure 20 – Structure hiérarchique de catégories du classifieur

5.1.2 Autres variables à tester

La densité, l'homogénéité des éléments de l'image, leur répartition, l'orientation, la couleur, l'unicité ou la division des éléments, la latéralisation, la sémantique, sont autant de variables à tester dans de futures expérimentations.

6 Conclusion

L'analyse de scènes complexe est un des domaines les plus difficiles et les plus étudiés en vision artificielle selon une approche complexe ascendante. Du pixel au groupe de pixels, du groupe de pixels aux traits, du trait aux objets, et des objets à la scène, les stratégies sont nombreuses et se basent toutes sur les techniques de traitements du signal. Il est important de noter que la stratégie utilisée par l'homme est inverse "coarse to fine". Ainsi le système visuel humain reconnaît la catégorie de la scène avant d'identifier les différents objets de l'image. Après analyse des résultats nous pouvons dire que l'homme utilise pour catégoriser des images naturelles complexes, des critères et des stratégies de différents niveaux. Le principal critère de catégorisation mis ici en valeur est la proportionnalité de l'élément, qui joue un rôle déterminant, tant au niveau de la classification, que du temps de réponse. Le filtrage joue également un rôle important dans l'activité de catégorisation, notamment dans le fait qu'il met en saillance les différents éléments de l'image. Le centrage permet d'augmenter ou de réduire l'ambiguïté en rendant saillant la partie centrale de l'image. A partir de ces résultats, nous sommes en train d'essayer de concevoir un robot classifieur qui utiliserait ces critères. Le but est qu'il soit assez généraliste et robuste. Pour revenir à la question initialement posée : "Selon

quelles critères une image fait partie d'une catégorie et pas d'une autre ?" Nous pouvons répondre que le critère de proportion joue un rôle essentiel, qu'un filtrage peut être utilisé avantageusement pour renforcer la saillance de certaines parties de l'images, et que le peut centrage permet de réduire l'ambiguïté.

La question d'origine était "comment catégorise t'on une image de scène complexe ? ". Nous pouvons dire que pour classifier, l'homme utilise différents processus, de bas niveau, hauts niveau et niveau intermédiaire. Il possède des critères qui sont propres aux éléments de l'image et à sa catégorie. L'homme dispose de très nombreuses catégories organisées en réseau. Il utilise des critères de catégorisation, pour juger de l'appartenance d'une image à une classe. De nombreux critères de niveau intermédiaire sont utilisés, la proportion n'en est qu'une illustration, il en a d'autres tel que la densité, l'homogénéité, l'orientation, la couleur.

Table des illustrations

BIBLIOGRAPHIE :

Ashby, W, Prinzmetal, A, Ivry, R & Maddox, G. (1996) A formal theory of feature binding object perception, *Psychological review*, 1996, **103**, pp. 165-192,

Beaudot, W. (1994). *Le traitement neuronal de l'information dans la rétine des vertébrés : Un creuset d'idées pour la vision artificielle*. Thèse de doctorat, Institut National Polytechnique, Grenoble.

Chauvin A., Guyader N., Marendaz C. & Hérault J. (2002). Argument for scene categorization with image amplitude spectra. *Perception,* **31**, 132 c.

Chauvin, A. (2003). *Perception des scènes naturelles : étude et simulation du rôle de l'amplitude, de la phase et de la saillance dans la catégorisation et l'exploration de scènes naturelles.* Thèse de doctorat, Université Joseph Fourier, Grenoble.

Chauvin, A., Hérault, J., Marendaz, C. & Peyrin, C. (2002). Natural scene perception: visual attractors and image neural computation and psychology. Dans W. Lowe et J. Bullinaria (Eds.), *Connexionist Models of Cognition and Perception*, World scientific press, 2002.

De Valois, R.L. & De Valois, K. (1988). Spatial Vision. *Oxford: Oxford University Press*.

De Valois, R.L., Albrecht, D.G. & Thorell, L.G. (1982a). Spatial Frequency selectivity of cells in macaque visual cortex. *Vision Research,* **22**, pp. 545-559.

De Valois, R.L., William, E., Hepler, Y. & Hepler, N. (1982b). The orientation and direction selectivity of cells in macaque visual cortex. *Vision Research,* **22**, pp. 531-544.

De Valois, R.L. & De Valois, K.K. (1990). *Spatial Vision.* Oxford University Press.

Derroche, S. & Ghorbel, F. (2004). Shape symmetry detection in gray-level using the analytical Fourier-Mellin representation. S*ignal processing,* **84**(1), pp. 25-39.

Fabre-Thorpe, M., Delorme, A., Marlot, C. & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience,* **13**, pp. *171-80.*

Guérin-Dugué, A. & Oliva, A. (2000). Classification of scene photographs from local orientations features. *Pattern Recognition Letters,* **21**, pp. 1135-1140.

Guyader, N. & Hérault, J. (2001 a). *Modélisation Biologique (Circuits rétiniens et Corticaux) pour la catégorisation d'Images.* Proceedings du IV Colloque Jeunes Chercheurs en Sciences Cognitives, Lyon, France.

Guyader, N. & Hérault, J. (2001 b). *Représentation espace-fréquence pour la catégorisation d'images.* Proceedings du GRETSI, Toulouse, France.

Guyader, N., Chauvin, A. & Le Borgne, H. (2002 (b)). Catégorisation de scènes naturelles: l'homme vs la machine. Actes *NSI 2002 : journées Neurosciences et Sciences de l'Ingénieur, La Londe-les-maures, France.*

Guyader, N., Le Borgne, H., Hérault, J. & Guérin-Dugué, A. (2002 (a)). *Towards the introduction of human perception in a natural scene classification system*. International workshop on Neural Network for Signal Processing (NNSP'2002), Martigny Valais, Suisse, September 4-6.

Le Borgne, H., Guyader, N., Guérin-Dugué, A. & Hérault, J. (2003). *Classification of images: ICA filters VS Human Perception.* Actes Seventh International Symposium on

Signal Processing and its Applications, vol 2, pp 251-254, July 1-4 2003, Paris, France.

Le Borgne, H. (2004). *Analyse de données par réseau de neurones auto-organisés*. Manuscrit de thèse, INPG, 1995.

Le Borgne, H., Guérin-Dugué, A. & Antoniadis A. (2004). Representation of images for classification with independent features. *Pattern Recognition Letters*, vol. 25, n°2, pp. 141-154.

Le Borgne, H., Guérin-Dugué, A. & Antoniadis, A. (2004). Representation of images for classification with independent features. *Pattern Recognition Letters,* **25**(2), pp. 141-154, january 2004.

Lipson, P., Grimson, E. & Sinha, P. (1997). Configuration based scene classification and image indexing. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Marendaz, C., Rousset, S. & Charnallet, A. (2003). Reconnaissance des scènes, des objets et des visages. *In A. Delorme & M. Fluckiger (Eds.), Perception et Réalité. Montréal : Gaëtan Morin Editeur.*

Marr, D. & Nishihara H.K. (1978). *Representation and recognition of the spatial organization of tree-dimendional shapes*. Proceeding of the Royal Society of London, B, 200, pp. 269-294.

Marr, D. (1982). "Vision: a computational investigation into the human representation and processing of visual information". *Freeman, San Francisco.*

Marshall, J. C. & Halligan, P. W. (1995). Seeing the forest but only half the trees? *Nature*, **373**(6514), pp. 521-523.

Mojsilovic, A. & Rogowitz, B. (2001). *Capturing image semantic with low-level descriptors*. Actes. International conference on image processing, vol 1, pp 18-21, Thessaloniki, Grèce, 7-10 octobre.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology,* **9**, pp. 353-383.

Navon, D. (1981). The forest revisited: More on global precedence. *Psychological Research,* **43**, pp. 1-32.

Neisser, U. (1964). Visual search. *Scientific American* 210(6), 94-102.

Oliva, A. (1995). *Perception de scènes: Traitement fréquentiel du signal visuel aspects psychophysiques et neurophysiologiques*. Thèse de Doctorat, Institut National Polytechnique, Grenoble.

Oliva, A. & Schyns, P.G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, **34**, pp. 72-107.

Oliva, A. & Torralba, A., Guérin-Dugué, A. & Hérault, J. (1999). *Global Semantic Classification of Scenes using Power Spectrum Templates*. Challenge for Image retrieval (CIR99), Newcastle, 1999.

Oliva, A. & Torralba, A. B. (2001). Modeling the Shape of the Scene: A holistic representation of the Spatial Envelope. *International Journal of Computer Vision,* **43**(3), pp. 145-175.

Oliva, A., Torralba, A. B., Guérin-Dugué, A. & Hérault, J. (1999). *Global semantic classification of scenes using power spectrum templates.* Proceedings of the Challenge of Image Retrieval (CIR99), Newcastle.

Rogowitz, B.E., Frese, T., Smith, J.R, Bouman, C.A. & Kalin, E. (1998). *Perceptual image similarity experiment.* SPIE Symposium on Electronic Imaging: Science and Technology, Conference on Human Vision and Electronic Imaging III, pp. 576-590.

Rosch, E. (1973), *On the internal structure of perceptual and semantic categories*, In T.E. Moore (Ed.),Cognitive Development and the Acquisition of Language, New York : Academic Press

Szummer, M. & Picard, R.W. (1998). *Indoor-outdoor image clasification*. IEEE international workshop on content-based access of image and video databases, Bombay, Inde, Janvier.

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

Thorpe, S.J., Bacon, N.M., Rousselet, G.A., Macé, M.C. & Fabre-Thorpe, M. (2002). Rapid categorisation of natural scenes: feedforward vs. feedback contribution evaluated by backward masking *Perception, **31*** suppl, pp. 150.

Torralba, A. & Oliva, A. (2003). Statistics of Natural Image Categories. *Network: Computation in Neural Systems*, **14**, pp. 391-412.

Treisman, A. Focused attention in the perception and retrieval of multidimensional stimuli. P*erception and Psychophysics*, 1977, 22, 1-11.

Treisman & Gelade (1980). *A Feature-Integration Theory of Attention,* NewYork

Vailaya, A., Jain, A. & Zhang H.J. (1998). On Image Classification: City vs. Landscape. *Pattern recognitions*, **31**(12), pp. 1921-1935.

Remerciements :
J'aimerais sincèrement remercier Emmanuelle Raynaud, Mohsen Ardabilian, Chen Liming, Jean-Claude Bougeant, Alain Pujol, Kostas Kosta, Olivier Koenig, Jérôme Moreau, Julien Chapuis, tous les participants et sujets, ainsi que Clotilde Roux, et Gabriel.