

---

# Indexation de documents multimédia par réseaux d'opérateurs

**Stéphane Ayache, Georges Quénot**

*Laboratoire d'Informatique de grenoble*

*385, rue de la Bibliothèque - B.P. 53 - 38041 Grenoble Cedex 9*

*Stephane.Ayache@imag.fr*

---

*RÉSUMÉ. Le franchissement du fossé sémantique entre les descriptions au niveau signal et au niveau sémantique est le principal problème à résoudre pour l'indexation multimédia. Les approches les plus avancées prennent en compte plusieurs types de descripteurs, plusieurs modalités et/ou le contexte pour améliorer la détection des concepts. Afin de maîtriser la complexité liée à l'intégration de données et de traitement hétérogènes que cela suppose, nous proposons une approche à base d'opérateurs organisés en réseaux flots de données. Un type unique de données, les numcepts, et un type unique d'unités de traitements, les opérateurs, sont utilisés à tous les niveaux dans ces réseaux. Cette approche offre une grande flexibilité pour la conception des systèmes d'indexation et elle facilite les expérimentations sur les architectures correspondantes. Nous décrivons une instance de ce modèle et plusieurs numcepts et opérateurs. L'approche a été testée sur les corpus TRECVID 2004 et 2005. Nous avons implémenté 7 réseaux et étudié l'influence de variations sur la précision moyenne des concepts détectés.*

*ABSTRACT. Bridging the semantic gap between signal level and semantic level descriptions is the main problem to solve for multimedia indexing. The most advanced approaches take into account several descriptor types, several modalities and/or the context to enhance concept detection. For mastering the complexity linked to the integration of the heterogeneous data and processes that this requires, we propose an approach based on operators organized into dataflow networks. A unique type of data, the numcepts, and a unique type of processing unit, the operators, are used at all levels in these networks. This approach offers a great flexibility for the design of indexing systems and it eases the experiments with the corresponding architectures. We describe an instance of this model and several numcepts and operators. We have tested the approach on TRECVID 2004 and 2005 corpora. We have implemented 7 of operators and studied the influence of variations on the accuracy of concept detection*

*MOTS-CLÉS : Indexation, Vidéo, Contexte, Fusion, Apprentissage.*

*KEYWORDS: Indexation, Video, Context, Fusion, Machine Learning.*

---

## 1. Introduction

L'indexation par le contenu est nécessaire pour une gestion automatisée de grands corpus de documents multimédia et pour fournir un moyen d'accès aux utilisateurs qui recherchent des informations multimédias. La majorité des travaux en indexation par le contenu de documents multimédia est liée au comblement du fossé sémantique qui sépare les éléments de bas niveau (signal) et les éléments de haut niveau (sémantique). Le document étant stocké sous forme numérique, on peut considérer qu'il n'est défini que par un ensemble de bits. L'assemblage de ces 0 et 1 compose un signal numérique, image ou vidéo. Comment passer de ce signal numérique à des descriptions sémantiques (ou conceptuelles) ? La majorité des approches existantes réduisent cet écart en recodant le signal numérique par un ensemble de descripteurs de bas niveau liés aux modalités image, audio et texte. Ensuite, un algorithme d'apprentissage supervisé est utilisé pour apprendre les régularités existantes entre les éléments de bas niveau et les descriptions sémantiques (Over *et al.*, 2005). Nous pensons qu'une telle approche n'est pas optimale et ne permet pas de combler efficacement le fossé sémantique. De nombreuses approches ont recours à des concepts intermédiaires pour franchir le fossé sémantique par parties (Naphade, 2004, Ayache *et al.*, 2006a, Snoek *et al.*, 2006, Lin *et al.*, 2002) selon une méthode d'apprentissage par ensemble, appelée « Stacking » (Wolpert, 1990). Par ailleurs, afin d'exploiter plusieurs sources d'informations hétérogènes dans les documents multimédias, de nombreux travaux s'intéressent à la fusion des modalités (Snoek *et al.*, 2005, Iyengar *et al.*, 2003, Naphade, 2004). Elle peut avoir lieu à plusieurs niveaux du processus d'indexation ; les approches les plus rencontrées sont appelées fusion « précoce » et « tardive ». Plus récemment, le contexte induit par l'usage de concepts intermédiaires et/ou de plusieurs sources d'informations fait aussi l'objet de travaux visant à combler le fossé sémantique pour une indexation sémantique de qualité (Naphade, 2004, Snoek *et al.*, 2006, Ayache *et al.*, 2006b).

Nous présentons ici une nouvelle approche basée sur la notion de réseaux d'opérateurs, pour combler le fossé sémantique par des éléments qui tendent à s'abstraire de plus en plus du contenu de bas niveau. Pour cela, nous proposons une homogénéisation des données manipulée pendant le processus d'indexation, nous les nommons *numcepts*, ils permettent de s'abstraire des modalités et d'envisager la fusion non pas au niveau modalité, mais au sein d'un continuum. Les numcepts sont mis en relation les uns avec les autres dans un réseau afin d'exploiter différentes formes de contexte et de permettre l'inférence ou la dérivation de nouveaux concepts. Plutôt que de considérer des différences qualitatives comme « niveau signal », « niveau intermédiaire », « niveau sémantique » ou autres encore, nous proposons d'ignorer toute différence qualitative de ce point de vue et de considérer celles-ci comme non pertinentes dans le cadre de notre problématique. De même, et dans le même esprit, nous proposons de ne considérer que des *opérateurs* (ou encore *modules*) prenant en entrée des numcepts et produisant en sortie d'autres numcepts. Ceci étant, pour clarifier notre approche, nous identifierons plusieurs catégories (ou types) de numcepts et plusieurs catégories (ou types) d'opérateurs sur ces numcepts.

Dans la section 2, nous présentons un modèle fondé sur les numcepts et les opérateurs. Dans la section 3, nous proposons une instance de cette approche et décrivons deux types de numcepts intermédiaires, appelés *percepts*, et plusieurs implémentations d'opérateurs sur ces numcepts. Nous validons le modèle proposé dans la section 4 où nous montrons 7 réseaux d'opérateurs mis en œuvre pour des expérimentations sur les corpus TRECVID 2004 et 2005 (Over *et al.*, 2005).

## 2. Numcepts et opérateurs

Les numcepts et les opérateurs sont les principaux constituants de l'approche proposée. Ils forment un réseau où les opérateurs sont des nœuds entre lesquels circule des flux de numcepts. Les opérateurs agissent dans le domaine des numcepts : un opérateur forme un nouveau numcept à partir d'un ou plusieurs numcepts. Soit  $\Phi$  le domaine des numcepts, un opérateur est une fonction  $f$  de  $\Phi^* \rightarrow \Phi$  :

$$\text{numcept}_x = f(\text{numcept}_{x_1}, \dots, \text{numcept}_{x_N})$$

La figure 1 représente un réseau pour la détection du concept DISCOURS DU PRÉSIDENT BILL CLINTON, les boîtes carrées symbolisent des opérateurs, les autres éléments sont des numcepts. Ils sont de nature différente mais leur point commun est d'établir une continuité entre les descriptions numériques et conceptuelles du document vidéo. Nous désignons donc par numcepts tout ce qui peut établir une continuité entre les éléments issus du signal et les éléments conceptuels qui décrivent une séquence vidéo. Il peut s'agir de simples octets, de pixels, de descripteurs de bas niveau, d'éléments étiquetés, etc. En sortie du système, des concepts tels que DISCOURS DU PRÉSIDENT AMÉRICAIN ou VOTE DU BUDGET FÉDÉRAL peuvent être reconnus.

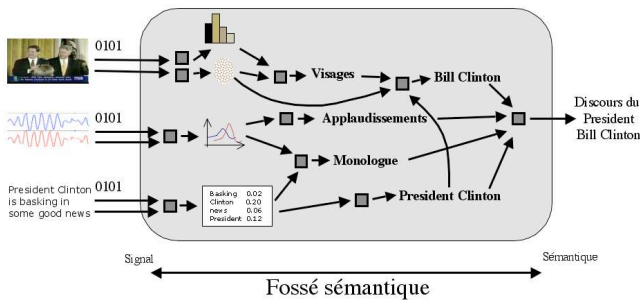


Figure 1. La fosse sémantique et le continuum de numcepts

Le système prend en entrée les valeurs numériques des pixels et/ou des échantillons sonores (par exemple) et produit en sortie des valeurs numériques (binaires ou analogiques). Celles-ci sont en général associées à des choses qui ont un sens pour les humains comme des concepts et des relations.

## 2.1. Les numcepts

Les *numcepts* ont pour vocation de clarifier, de généraliser et d'unifier un certain nombre de notions intervenant dans les différents traitements de l'information entre le niveau numérique (ou signal) et le niveau conceptuel (ou sémantique). On y rencontre en effet toutes sortes d'objets comme du signal, des pixels, des échantillons, des descripteurs, des chaînes de caractères, des caractéristiques, des contours, des régions, des blobs (Carson *et al.*, 2002), des points d'intérêts, des formes, des textures, des vecteurs de mouvement, des concepts intermédiaires, des percepts, des sujets, des thèmes, des concepts, des relations, etc. Ceci est amplifié par les approches en couches ou en réseau (présentement les plus performantes, *stacking* par exemple (Wolpert, 1990)) qui font parfois intervenir des *entités intermédiaires* qui ne sont plus clairement ni des descripteurs numériques au sens classique ni des concepts au sens classique également (c'est à dire quelque chose qui a un sens pour un être humain).

Le mot *numcept* est dérivé des mots *number* (ou description numérique) et *concept* (ou description conceptuelle) et vise à décrire quelque chose qui généralise et unifie ces deux notions a priori qualitativement différentes. En effet, une des difficultés pour combler le fossé sémantique tient justement à ce saut qualitatif et à cette différence de nature que l'on peut ressentir intuitivement entre ces deux types d'information ou de niveaux, traditionnellement appelés *niveau signal* et *niveau sémantique*. Du point de vue traitement de l'information, une telle différence qualitative n'existe pas. Tous les éléments considérés, quel que soit leur niveau d'abstraction, sont représentés sous une forme numérique. C'est seulement l'interprétation qu'un humain donnera à ces différents éléments qui produira, éventuellement, une différence qualitative entre eux. Bien entendu, on reconnaîtra toujours comme numériques des pixels d'une image ou des échantillons sonores d'un segment audio et on reconnaîtra toujours comme concepts les éléments présents à l'autre bout de la chaîne de traitements comme la production d'une ou plusieurs étiquettes (ou l'attribution de valeurs binaires ou continues à celles-ci). Cette définition englobe donc les concepts : les concepts et les descriptions numériques sont des numcepts proches des extrémités du continuum.

En faisant ce genre d'unification, nous sommes passés d'une approche hétérogène à une approche homogène et nous nous sommes également affranchis des rigidités des approches en couches selon des schémas prédéfinis (par exemple classant les traitements en bas, moyen et haut niveau). Cette façon de voir les choses ne correspond pas à un changement radical mais elle offre plus de souplesse et de liberté dans la conception et la mise en œuvre des systèmes d'indexation par concepts. Elle permet d'envisager des architectures riches et variées sans avoir à se poser la question du type de données manipulées ou du type d'opérateur utilisé. Toutes les combinaisons de types de données et d'opérateurs deviennent possibles et peuvent être explorées expérimentalement.

### 2.1.1. Catégories de numcepts

Bien que nous considérons les numcepts comme des éléments homogènes et de nature continue entre les descriptions numériques et conceptuelles, nous donnons quelques exemples de catégories pour préciser notre approche. Nous décrivons deux catégories des numcepts, les *numcepts de bas niveau* sont typiquement extraits au début du processus d'indexation, ils se situent près de la source du réseau. Les *numcepts sémantiques* ont la particularité d'être compréhensible par un être humain, ils peuvent être extrait (avec plus ou moins de précision) à tout moment du processus d'indexation, mais sont typiquement des numcepts qu'on trouve en sortie du réseau d'opérateurs.

– Les *numcepts de bas niveau* caractérisent l'ensemble des séquences vidéo, et sont, par définition, dépendant des modalités à partir des quelles ils sont extraient. Ils correspondent aux descripteurs de bas niveau d'une portion du document vidéo. Pour la modalité visuelle, il peut s'agir de simples pixels en niveau de gris (utilisés notamment pour la détection de visages), d'histogrammes de couleurs locaux ou globaux, de descripteurs de textures ou de mouvements. Un numcept de bas niveau est caractérisé par un ensemble d'éléments décrits, associés aux algorithmes d'extractions utilisés. Les attributs numériques associés aux numcepts primitifs sont typiquement des descripteurs de bas niveau.

– Les *numcepts sémantiques* sont des numcepts associés à une étiquette sémantique. Nous les considérons comme des concepts. Leur formation nécessite l'usage d'un algorithme de classification via un apprentissage supervisé. La classification est basée sur un ensemble de numcepts décrivant l'élément vidéo considéré. Similairement aux numcepts non étiquetés, cette catégorie de numcepts est caractérisée par un type d'élément vidéo, le modèle correspondant à la classe sémantique, et le contexte du numcept. Sa représentation numérique est issue de la phase de classification, c'est une valeur réelle correspondant à la valeur de prédiction de la classe sémantique décrite par le concept.

Tous les numcepts ne sont pas supposés appartenir à l'une ces catégories ; nous supposons au contraire une continuité entre elles. Certains numcepts peuvent être générés par des algorithmes de « clustering », et d'autres peuvent être le produit d'une combinaison de numcepts. Dans tous les cas, les numcepts ont tous en commun le fait d'être liés à un ensemble d'attributs numériques.

## 2.2. Les opérateurs

Les opérateurs sont destinés à combler le fossé sémantique par transformation d'un certain nombre de numcepts en d'autres numcepts plus abstraits. Pour y parvenir, un opérateur peut nécessiter l'usage d'une phase d'apprentissage pour optimiser une transformation précise. Nous distinguerons et introduisons quatre types d'opérateurs : les opérateurs d'extraction, de classification, de fusion et de contexte. Un opérateur d'extraction vise à la formation de numcepts de bas niveau par sélection ou combi-

naissance des attributs numériques des numcepts entrants. Un opérateur de classification permet de former des groupes (ou classes) d'éléments vidéo, ils peuvent être étiquetés ou non. Les opérateurs de contexte tirent parti d'informations contextuelles pour former de nouveaux numcepts. Nous envisageons plusieurs formes de contexte : sémantique, spatial et temporel. Enfin, les opérateurs de fusion permettent de combiner plusieurs numcepts pour en former de nouveaux.

- Un *opérateur d'extraction* prend en charge la phase d'échantillonnage nécessaire à la formation d'éléments vidéo, puis en donne une représentation de bas niveau. Typiquement, ce processus vise à réduire la quantité d'informations contenue dans le signal de façon à ne retenir que celles qui sont pertinentes à une tâche donnée. Cela revient à extraire une représentation de l'élément vidéo par des caractéristiques plus simples, comme des couleurs, des lignes, des mouvements, des fréquences, etc. Selon notre formalisme, un opérateur d'extraction prend en entrée des numcepts correspondant à du signal, et crée en sortie des numcepts de bas niveau.

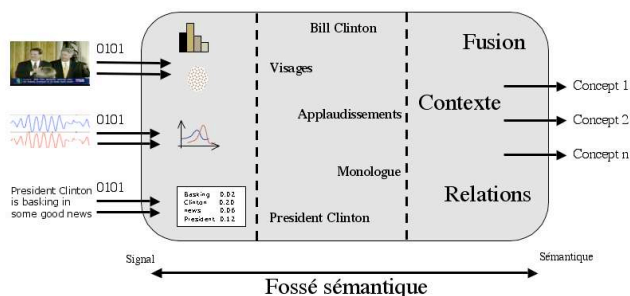
- Un *opérateur de classification* vise à identifier des groupes d'éléments vidéo qui partagent des caractéristiques communes. D'après notre formalisme, ces groupes sont similaires aux numcepts non-étiquetés ou sémantiques. L'algorithme utilisé peut avoir recours à une phase d'apprentissage, qui doit nécessairement être supervisée pour aboutir à la formation de groupes étiquetés (numcepts sémantiques). Les éléments considérés par l'algorithme de classification correspondent aux numcepts entrant dans l'opérateur.

- Un *opérateur de fusion* réalise une combinaison de numcepts de natures quelconques. La fusion de numcepts considérés dans le cadre de notre approche vise à combiner des numcepts formés à tout niveau du processus d'indexation pour former des numcepts sémantiques. Un opérateur de fusion peut combiner des numcepts issus de différentes modalités du document vidéo, aux propriétés spatio-temporelles différentes, associés à des éléments vidéo variés (local ou global) et/ou combiner des numcepts de bas niveau avec des numcepts sémantiques. La fusion peut être soit définie par des règles explicitement fixées selon des hypothèses a priori soit par optimisation d'une fonction de combinaison, nécessitant alors l'usage d'une phase d'apprentissage.

- Un *opérateur de contexte* met en relation plusieurs numcepts afin d'exploiter les circonstances et conditions dans lesquelles un numcept apparaît. Dans notre étude, nous considérerons que le contexte participe à la formation de nouveaux numcepts en exploitant la configuration d'autres numcepts présents dans la séquence vidéo traitée. La configuration de ces numcepts est définie par leurs valeurs numériques mais aussi leurs propriétés spatio-temporelles : où et quand les numcepts sont activés. En d'autres termes, les dispositions spatio-temporelles de certains numcepts forment le contexte d'un numcept donné. Le contexte fournit alors ses contraintes pour la prise de décision sur un numcept. Nous décrirons dans la partie 3.2 deux formes de contexte : les contextes sémantique, et spatial.

### 3. Mise en œuvre : une instance du continuum

Nous avons décrit un modèle générique comportant des notions telles que numcepts, opérateurs et contextes qui s'agencent en réseau afin de réaliser une continuité entre des descripteurs de niveaux signal et sémantique des éléments vidéo. Dans cette partie, nous présentons une instance de ce modèle : une réelle continuité étant impossible en pratique (il faudrait une infinité de numcepts et opérateurs), nous faisons le choix de discrétiser le continuum et proposons une indexation sémantique via différents types de numcepts pour combler le fossé sémantique par parties. La figure 2 illustre notre choix d'instanciation du modèle : le continuum est découpé en trois parties dans lesquelles différents types de numcepts et opérateurs sont mis en œuvre.



**Figure 2.** Instanciation du modèle

La première partie concerne l'extraction et la représentation des éléments vidéo. Cette étape aboutit à la formation de numcepts de types descripteurs de bas niveau. L'opérateur ayant cette responsabilité est défini de manière à extraire un élément vidéo (séquence, image, région d'image, échantillon, etc.) depuis le signal et à le représenter par le descripteur de bas niveau visé.

La deuxième partie vise à s'affranchir des modalités du document vidéo. Les numcepts sont des éléments intermédiaires qui se placent entre les couches signal et sémantique de la phase d'indexation. Nous décrivons deux classes de numcepts intermédiaires, ceux dérivés des modalités visuelles et ceux dérivés des textuelles. Nous les appelons *percepts* par analogie avec le système cognitif humain, car ils sont issus d'une phase de perception (opérateurs d'extraction + numcepts de bas niveau) et permettent la formation de concepts.

La troisième partie concerne les aspects de contexte et de fusion (ou combinaison des numcepts). Nous décrivons dans cette étape des opérateurs de contexte, ainsi que des opérateurs de fusion. Les numcepts formés lors des étapes précédentes sont mis en relation pour dériver les concepts.

### 3.1. Les numcepts intermédiaires

Nous introduisons dans cette partie la notion de numcepts intermédiaires, ils s'appuient sur les numcepts de bas niveau et sont à l'origine de la formation de numcepts plus abstraits. Ici, nous décrivons deux types de numcepts intermédiaires, en particulier, nous développerons l'idée que ces éléments constituent un contexte utile pour dériver de nouveaux concepts. Ils agissent tels des indices pour favoriser (ou non) le processus de décision de numcepts plus sémantiques.

Afin de faciliter la prise de décision et éviter le problème lié aux espaces de grandes dimensions, les numcepts intermédiaires permettent de réduire la taille des données numériques associés aux numcepts de bas niveau. Nous avons fait le choix de mettre en œuvre des numcepts sémantiques, plus directement intelligibles et manipulables dans le cadre des phases de développement et de test dont ces travaux ont fait l'objet. Outre le fait que ces éléments nous ont permis d'avoir un retour direct sur la qualité des descripteurs de bas niveau et de l'apprentissage, le recours à des numcepts sémantiques a permis d'alimenter notre base de concepts. Ils sont particulièrement utiles pour une tâche de recherche de documents multimédia. En revanche, ce choix induit la nécessité de passer par une phase d'apprentissage supervisé, impliquant de constituer une base d'exemples positifs et négatifs pour chacun des éléments sémantiques visés. Pour cela, nous avons utilisé des ressources existantes fournies notamment par les campagnes d'évaluations TRECVID.

Dans ce qui suit, nous décrivons deux classes de numcepts intermédiaires, nous les appelons *percepts visuels* et *percepts textuels*. Ils sont dans les deux cas rattachés à des éléments vidéo locaux (respectivement blocs d'images et segments audio), et visent à enrichir la description d'éléments vidéo de granularité plus large.

#### 3.1.1. *Percepts visuels*

Les *percepts visuels* permettent de s'abstraire du contenu numérique lié aux descripteurs de bas niveau extraient dans la modalité visuelle. Au niveau local, les numcepts de bas niveau décrivent la modalité visuelle pour chaque bloc des images clés du plan. En considérant  $N$  descripteurs par bloc et  $B$  blocs par image, chaque image est décrite par  $N \times B$  valeurs numériques. Afin de faciliter la prise de décision sur le contenu sémantique d'une séquence vidéo, notre approche consiste à associer un ensemble de *percepts visuels* par bloc d'image. Les descripteurs de bas niveau sont sujets à des variations liées à la qualité du signal visuel, la qualité du processus d'extraction et la difficulté d'obtenir des descripteurs invariants au contenu sémantique de l'image. Pour s'affranchir de ces variations, les *percepts visuels* sont issus d'un processus d'apprentissage visant à regrouper des blocs d'images partageant un contenu sémantique commun. Une région correspondant à de la végétation peut par exemple être caractérisée au niveau signal de multiples façons selon le type et l'état du végétal (texture et couleur différentes). Une caractérisation sémantique des régions contenant des végétaux est alors invariante à ces variations, ce qui assure une description plus robuste et invariante du média.



Nous avons formé une base de 15 percepts visuels, il s'agit d'éléments sémantiques dont l'identification peut être faite sans ou avec peu de contexte : nous considérons alors que les descripteurs de bas niveau d'un seul bloc d'image sont suffisant pour détecter de tels éléments sémantiques (avec des taux d'erreur acceptables). Nous avons défini 15 éléments sémantiques qui sont d'une part aisément identifiables dans des blocs de tailles  $32 \times 32$  pixels, et d'autre part suffisamment variés pour être réutilisés comme descripteurs d'éléments vidéo de granularité moins fines (cf. contexte). Voici quelques exemples de percepts visuels : FEU, VÉGÉTATION, EAU, PEAU et CIEL.

### 3.1.2. *Percepts de sujets*

Les percepts de sujets sont dérivés de la modalité textuelle, ils visent à identifier des sujets larges (ou thèmes) abordés dans une séquence vidéo à partir de ce qui est prononcé. Les sujets détectés forment une base d'indices sémantiques pour participer à la caractérisation des séquences vidéo. Le panel de sujets susceptibles de participer à une caractérisation sémantique pertinente des séquences vidéo dépend du type de corpus traité. Nous avons utilisé le corpus Reuters (RCV1) (Lewis *et al.*, 2004), qui contient plus de 800000 dépêches annotées, établies entre 1996 et 1997. Les catégories Reuters sont organisées en hiérarchie, il y a 25 catégories mères (telles que ECONOMIE, POLITIQUE, SPORTS, GUERRE, ...) qui se spécialisent en sous-catégories selon un à trois niveaux de profondeur. L'ensemble des nœuds (pères, intermédiaires et feuilles) totalise 103 catégories.

Pour classer les segments audio du corpus TRECVID selon les catégories Reuters, nous nous sommes basés sur le classifieur de Rocchio, issu du modèle vectoriel. Ce classifieur a été largement utilisé dans la communauté de la catégorisation textuelle, il se distingue par sa rapidité d'apprentissage et de classification. Pendant la phase d'apprentissage, un classifieur de Rocchio crée un vecteur prototype pour chaque classe (les catégories Reuters) puis, pendant la phase de classification, chaque segment audio du corpus vidéo est comparé aux prototypes via une distance normalisée (de type cosinus). A la fin du processus, plusieurs percepts sont associés à chaque segment audio, chaque percept contient une valeur numérique correspondant au score de reconnaissance d'une catégorie Reuters.

## 3.2. *Opérateurs de contexte*

Nous avons identifié deux formes de contexte : les contextes sémantique et spatial, le premier concerne les numcepts issus d'un opérateur de classification, tandis que le deuxième exploite l'agencement spatial des percepts visuels.

### 3.2.1. *Contexte sémantique*

Le contexte sémantique met en relation plusieurs numcepts générés par un opérateur de classification. L'opérateur de contexte consiste à créer un nouveau numcept dont la structure numérique sous-jacente contient le score de classification des différents numcepts. Nous mettons en évidence ces relations en alignant les scores de

classification dans un vecteur. L'alignement conserve alors la sémantique de chaque numcept : la  $n^{ieme}$  composante est affectée au  $n^{ieme}$  numcept.



**Figure 3.** *Scènes de studio*

Prenons plusieurs exemples de scènes de studio (figure 3), ceux sont des scènes qui contiennent essentiellement un décor en arrière plan, un présentateur, et pas de végétation (par exemple). Nous dirons que le contexte sémantique du numcept SCÈNE DE STUDIO est constitué des percepts DÉCOR DE STUDIO, PEAU et VÉGÉTATION. La moyenne des scores de classification de ces trois percepts obtenus sur les images de la figure 3 sont respectivement 88%, 86% et 2%. Ces scores ont été obtenus via un classifieur exploitant le contexte spatial, présenté dans la section suivante. L'opérateur de contexte sémantique aligne ces trois scores dans une structure numérique qui constituera le nouveau numcept. Pour inférer le numcept sémantique SCÈNE DE STUDIO à partir de ce dernier, il est nécessaire de passer par un opérateur de classification.

### 3.2.2. Contexte spatial

L'agencement des numcepts d'origine visuelle est particulièrement discriminant pour la catégorisation de scènes. Dans la figure 3, il apparaît que l'agencement topologique des régions contenant du décor de studio est très caractéristique de ce type de scène. Typiquement, le présentateur est soit au milieu, soit sur un côté et dans tous les cas le reste de l'image contient du décor. Une façon de représenter l'agencement topologique des régions consiste à créer une structure dans laquelle les attributs numériques des percepts visuels sont alignés selon l'ordre des blocs (du haut à gauche vers le bas à droite). Le nouveau numcept est décrit par un vecteur de dimension égale au nombre de percepts visuels entrant. L'alignement de ces valeurs dans un vecteur permet de conserver l'information relative au positionnement spatial de chaque bloc : la première composante correspond au bloc du coin supérieur gauche de l'image, la deuxième correspond au bloc qui suit, et ainsi de suite jusqu'au bloc du coin inférieur droit de l'image.

### 3.3. Opérateur de fusion (ou combinaison de numcepts)

Les approches classiques de fusion dites « précoce » et « tardive » visent essentiellement à réaliser l'étape cruciale qu'est la fusion des modalités, le plus souvent via un algorithme de classification. Une fusion précoce consiste à combiner les descripteurs de bas niveau issus de différentes modalités avant d'effectuer la classification, tandis qu'une fusion tardive combine les scores de classification obtenus sur chacune des

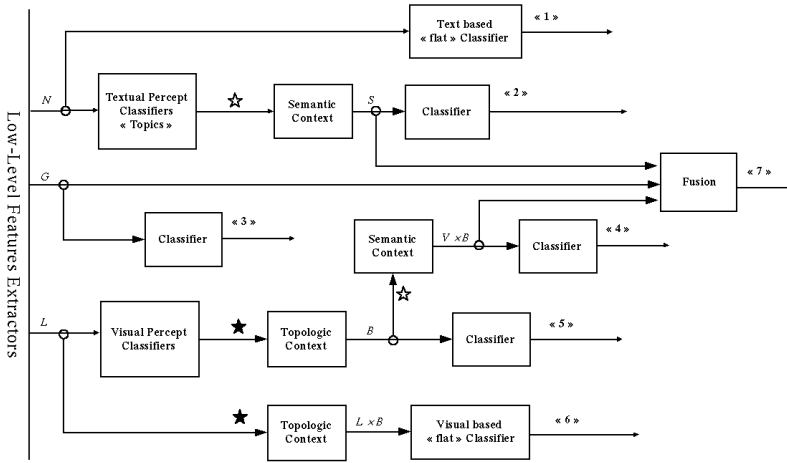
modalités par un deuxième niveau de classifieur. Ils opèrent réciproquement une fusion de numcepts de bas niveau ou une fusion de numcepts sémantiques. Ces schémas sont faciles à mettre en œuvre et sont majoritairement utilisés dans la communauté d'indexation de documents multimédia. Un processus de fusion doit être en mesure d'extraire une information de meilleure qualité (en termes de précision) que celles obtenues par le biais d'une seule des sources utilisée. Les schémas de fusion classiques sont cependant limités lorsqu'il s'agit de combiner des données hétérogènes, l'hétérogénéité venant soit d'une grande différence de dimension entre les vecteurs issus des différentes sources, soit de performances très différentes selon les sources. Ce type de situation est particulièrement fréquent en indexation de documents multimédia étant donné la disparité des descripteurs envisageables. Suivant notre approche, la fusion tire parti de l'homogénéisation faite sur les numcepts : un opérateur de fusion est alors en mesure de combiner toute sorte de numcepts, quelque soit leur origine. Plusieurs opérateurs de fusion à base de classifieurs SVM sont décrits précisément dans (Ayache *et al.*, 2007, Ayache *et al.*, 2006a).

#### 4. Validation

Afin d'évaluer l'apport de l'architecture proposée, nous comparons plusieurs réseaux d'opérateurs. Ils visent tous à extraire un ensemble de concepts dans les séquences vidéo et partagent des parties communes (sous réseaux communs). Les réseaux considérés permettent d'étudier différents agencements d'opérateurs et de numcepts. En particulier, nous étudions l'apport d'une indexation via les percepts de sujets et visuels ainsi que l'impact des opérateurs de contexte sur la qualité (en précision moyenne) de détection des concepts.

Ces expérimentations ont été conduites selon le protocole proposé par les campagnes d'évaluation TRECVID. Elles fournissent un corpus de documents vidéo (des journaux télévisés de plusieurs pays), une segmentation en plans de chaque document, le texte issu d'un processus de transcription automatique de la parole, ainsi qu'un ensemble de séquences annotées par concept pour entraîner les algorithmes d'apprentissage. Les résultats ont été générés avec l'outil `trec_eval`, la mesure de performance est la précision moyenne obtenue sur tous les taux de rappels (MAP). La figure 4 représente une vue globale des réseaux mis en œuvre. Pour ne pas surcharger la figure, seuls les opérateurs intervenant après la phase d'extraction des descripteurs sont représentés. Les sept flux de numcepts sortant (numérotés de 1 à 7) correspondent aux résultats d'un réseau. Ils contiennent les scores de détection des concepts TRECVID.

Les lettres indiquées sur la figure désignent le nombre d'attributs numériques des numcepts circulant dans les flux :  $N$ ,  $G$  et  $L$  correspondent respectivement à la dimension des descripteurs de bas niveau des modalités textuelle, visuelle globale et visuelle locale (voir section 4.1).  $B$  correspond au nombre de blocs considérés pour les descripteurs locaux, et les lettres  $S$  et  $V$  désignent respectivement le nombre de Percepts de Sujets et de Percepts Visuels. Enfin, les étoiles décorent les flux entrant dans les opérateurs de contexte : les étoiles blanches sont associées aux opérateurs de contexte

**Figure 4.** Réseau

sémantique et signifient que le flux correspondant contient plusieurs numcepts sémantiques. Les étoiles noires sont associées aux opérateurs de contexte spatial, le flux correspondant contient les numcepts issus de chaque bloc d'une image. Les opérateurs de classification de percepts, qui figurent en gras sur le schéma, sont déployés en parallèle pour classer les différents percepts. Chacun forme un numcept sémantique (percept) dont le seul attribut numérique est un score de classification. Les autres opérateurs de classification sont des classifieurs SVM à noyau Gaussien (Chang *et al.*, 2001) dont les paramètres sont optimisés par validation croisée pour chacun des concepts. L'opérateur de fusion combine trois flux de numcepts selon une combinaison des schémas de fusion précoce et tardive, décrite précisément dans (Ayache *et al.*, 2006a).

#### 4.1. Numcepts de bas niveau

L'objectif des expérimentations étant principalement d'évaluer différents réseaux d'opérateurs, les descripteurs de bas niveau considérés lors des expérimentations n'ont pas fait l'objet d'une étude approfondie, ceux sont néanmoins des descripteurs de bas niveau classiquement utilisés en indexation multimédia. Pour la modalité textuelle, nous considérons que les descripteurs de bas niveau sont les termes issus d'un opérateur de transcription automatique de la parole. Après suppression des termes vides ("the", "a", "is" ...), il reste environ 18000 termes dans le corpus TRECVID 2005. Parmi ceux là, 12500 apparaissent au maximum cinq fois. Finalement, après extraction des racines des mots via l'algorithme de Porter, il reste environ  $N = 12000$

termes. Dans la modalité visuelle, nous considérons des descripteurs globaux et locaux dans les images clés. La description locale est basée sur un découpage en  $B = 260$  ( $20 \times 13$ ) blocs semi-recouvrants de taille  $32 \times 32$  pixels dans lesquels nous avons extrait des descripteurs de couleur, texture et mouvement.

- Descripteurs Locaux : Moments de couleurs dans l'espace RGB (3 moyennes + 6 covariances); Filtres de Gabor (3 échelles  $\times$  8 orientations); 2 coordonnées des blocs ( $G = 35$ ).

- Descripteurs Globaux : Histogramme 3D ( $4 \times 4 \times 4$ ) dans l'espace RGB; Filtres de Gabor (5 échelles  $\times$  8 orientations); Moments des vecteurs de mouvements obtenus par calcul du flot optique (2 moyennes + 3 covariances) ( $L = 109$ )

## 4.2. Expérimentations

Nous présentons deux séries d'expérimentations conduites sur les corpus TRECVID 2004 et 2005. Nous commençons par présenter une étude préliminaire conduite afin de quantifier l'apport d'une indexation par numcepts intermédiaires (les percepts) sur la modalité visuelle. La deuxième série d'expérimentation a été conduite sur le corpus TRECVID 2005 où plusieurs opérateurs de contexte, de classification et de fusion ont été déployés pour une indexation multimodale des documents vidéo.

### 4.2.1. Réseaux pour l'indexation d'images fixes

Cette étude a été conduite sur un sous-ensemble d'apprentissage du corpus TRECVID 2004 contenant une annotation locale de 48810 images clés, la moitié pour l'apprentissage et l'autre pour le test. Nous avons considéré cinq concepts qui sont facilement détectables dans des blocs de  $32 \times 32$  pixels et qui sont liés par des relations sémantiques et spatiales. Ces 5 concepts sont aussi utilisés au niveau local : ce sont 5 des 15 percepts visuels considérés. Chaque percept a été entraîné à partir de 2048 blocs positifs et le double de blocs négatifs, tirés aléatoirement. Nous avons utilisé, en moyenne, 500 images positives par concept (et le double de négatives) pour la classification des images globales. Trois réseaux sont proposés (numérotés de 3 à 5). Dans le réseau 5, les numcepts de bas niveau de chaque bloc alimentent directement un opérateur topologique, puis un classifieur « à plat » classe les concepts dans un espace à  $L \times B = 9100$  dimensions. Le réseau 4 introduit un opérateur supplémentaire après les opérateurs d'extraction pour la classification des percepts visuels. Cette étape induit une réduction de dimensions de l'espace des caractéristiques dans lequel les concepts sont classés : seules  $B = 260$  dimensions agissent sur la classification. Le réseau 3 se base aussi sur les percepts visuels, deux opérateurs de contexte sont mis en série avant l'opérateur de classification : les taux de classification des 5 percepts sur tous les blocs forment un espace caractéristiques de  $B \times V = 1300$  dimensions.

Le tableau 4.2.1 montre la performance relative des trois réseaux pour la détection des concepts sur l'image globale (en MAP), et l'impact de chacun sur les temps d'apprentissage (en mn/CPU). Les temps contiennent à la fois le temps nécessaire pour

	Numcepts Visuels (5) Bas Niveau Contexte Spatial	Percepts Visuels (4) Contexte Spatial	Percepts Visuels (3) Contexte Spatial + Sémantique
Building	0,1927	0,3077	0,4230
Sky	0,5453	0,4331	0,5606
Studio Setting	0,8905	0,7675	0,9106
Greenery	0,4623	0,7207	0,7283
Skin	0,3421	0,4562	0,4280
All	0,4866	0,5370	0,6101
Training Time	836	418	484

**Tableau 1.** *Résultats de la première série d'expérimentation (sur TRECVID 2004)*

l'apprentissage des percepts (sur les blocs) et le temps d'apprentissage des concepts selon le type de contexte considéré, en prenant en compte la validation croisée. Pour cela, nous avons mis en parallèle 11 processeurs Pentium4 à 3Ghz.

Le réseau 5 correspondant à l'usage des numcepts de bas niveau nécessite deux fois plus de temps d'apprentissage que les deux autres réseaux, et ce pour une performance en précision moyenne inférieure. Le réseau 4 montre que l'apport des percepts visuels apporte un gain sur les ressources processeurs utilisées, certainement dû à la réduction de dimension induite par les numcepts intermédiaires. Par ailleurs, le gain qualitatif est lié à l'usage d'un opérateur de classification supplémentaire, qui vise à réduire la variabilité des attributs numériques entrant dans le classifieur final. Dans le réseau 3, l'ajout d'un opérateur de contexte sémantique améliore encore la précision moyenne de 13 % comparé au réseau 4, le classifieur final tire parti des relations sémantiques implicites qui lient les cinq concepts.

#### 4.2.2. Réseaux pour l'indexation de séquences vidéo

La deuxième série d'expérimentation a été conduite sur le corpus TRECVID 2005 contenant environ 80 heures de journaux télévisés (150000 images clés) provenant de chaînes télévisées anglophones, libanaises et chinoises. Les deux modalités considérées sont le texte et l'image, le texte issu de chaque segment audio (identifié par le processus de transcription de la parole) est aligné sur les images clés. Nous pouvons alors considérer conjointement les modalités visuelles et textuelles pour la classification des séquences vidéo. Cette évaluation porte sur les 10 concepts de la campagne TRECVID 2005, pour cela nous avons utilisé les 103 percepts de sujets et 15 percepts visuels décrits dans la section 3.1.

Le réseau 1 classe directement les concepts à partir des  $N = 12000$  termes, chacun décrit par une valeur booléenne pour sa présence ou non dans la séquence vidéo considérée. Le réseau 2 introduit l'usage des percepts de sujets, ils sont mis en relation par un opérateur de contexte sémantique. Le réseau 3 consiste à classer directement les concepts à partir des  $G = 109$  descripteurs visuels globaux. Le réseau 4 est équivalent

à celui décrit précédemment et le réseau 7 introduit l'usage d'un opérateur de fusion : les 103+1300 percepts sont combinés avec les  $G$  numcepts de bas niveau. L'opérateur de fusion est décrit dans (Ayache *et al.*, 2006a).

	(1)	(2)	(3)	(4)	(7)
People Walk./Run.	0,0046	0,0148	0,0686	0,1447	0,1286
Explosion/Fire	0,0004	0,0127	0,0108	0,0406	0,0453
Map	0,0016	0,0318	0,0476	0,1724	0,1923
US Flag	0,0022	0,0073	0,0412	0,0603	0,0736
Building	0,0114	0,0070	0,0755	0,2402	0,2612
Waterscape	0,0013	0,0034	0,0800	0,1899	0,1876
Mountain	0,0011	0,0054	0,1086	0,2857	0,2931
Prisoner	0,0008	0,0044	0,0002	0,0002	0,0002
Sports	0,0007	0,0322	0,0958	0,2682	0,3018
Car	0,0023	0,0082	0,0834	0,1811	0,2025
All	0,0026	0,0127	0,0612	0,1583	0,1686

**Tableau 2.** Résultats de la deuxième série d'expérimentation (sur TRECVID 2005)

Le tableau 4.2.2 montre les précisions moyennes obtenues pour la classification des 10 concepts via les 5 réseaux considérés. Il apparaît que la classification basée sur la modalité textuelle donne des résultats bien inférieurs à ceux obtenus par la modalité visuelle. Cela est certainement dû à la faible qualité des transcriptions de la parole obtenues sur des vidéos en arabe et chinois puis traduits automatiquement en anglais. Néanmoins, en comparant les réseaux 1 et 2, nous relevons une amélioration significative des précisions moyennes obtenues par l'usage des percepts de sujets. La comparaison des réseaux 3 et 4 montre que le gain en précision moyenne obtenu par l'usage des percepts visuels comparés à des descripteurs globaux d'images est très significatif. Enfin, le réseau 7 montre que malgré les grandes différences qualitatives obtenues par les réseaux 2, 3 et 4, une combinaison des numcepts intervenant dans ceux-ci améliore la précision moyenne des 10 concepts.

A titre de comparaison, le meilleur système à TRECVID 2005 (IBM) a atteint une performance de 0.33 et la performance médiane était de 0.16. Certains de nos choix ne sont sans doute pas optimaux, notamment en ce qui concerne la quantité et la nature exacte de nos descripteurs de bas niveau mais nous pensons que l'approche a tout de même été validée et que la plupart de nos conclusions par exemple sur l'apport de la multi-modalité et du contexte resteraient valables avec plus de descripteurs ou de meilleurs descripteurs.

## 5. Conclusion

Nous avons présenté une architecture basée sur des réseaux d'opérateurs flots de données pour l'indexation par concepts de documents multimédia. Plusieurs opérateurs ont été implémentés et visent chacun à combler une partie du fossé sémantique.

Nous avons également introduit la notion de numcept afin d'unifier les entités circulant dans le réseau. Le modèle a l'avantage d'être souple et de permettre de s'affranchir des barrières communément tracées entre les différents niveaux de données traitées et entre les différents types de traitement impliqués. De nombreuses variantes de réseaux peuvent ainsi être envisagées pour améliorer la détection de concepts dans les documents multimédias. Nous avons validé l'approche par plusieurs expérimentations conduites sur les corpus TRECVID et montré que l'usage de plusieurs opérateurs et numcepts permettent une indexation sémantique plus précise. Ces travaux vont être poursuivis avec la génération et l'évaluation automatique de réseaux d'opérateurs afin de rechercher le ou les réseaux susceptibles de franchir le fossé sémantique le plus efficacement possible en termes de précision du résultat et/ou de coût de calcul.

## 6. Bibliographie

- Ayache S., Quénot G., Gensel J., « Classifier Fusion for SVM-Based Multimedia Semantic Indexing », *In proceedings of ECIR*, 2007.
- Ayache S., Quénot G., Gensel J., Satoh S., « Using Topic Concepts For Semantic Video Shots Classification », *In proceedings of CIVR*, 2006a.
- Ayache S., Quénot G., Satoh S., « Context Based Conceptual Image Indexing », *In proceedings of ICASSP*, 2006b.
- Carson C., Belongie S., Greenspan H., Malik J., « Blobworld : Image Segmentation Using Expectation-Maximization and Its Application to Image Querying », *Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, n° 8, p. 1026-1038, August, 2002.
- Chang C.-C., proceedings of C.-J. L., *LIBSVM : A Library for Support Vector Machines*. 2001, Software available at <http://www.csie.ntu.edu.tw/~cjlln> proceedings of libsvm.
- Iyengar G., Nock H., « Discriminative Model Fusion for Semantic Concept Detection and Annotation in Video », *In proceedings of ACM Multimedia*, 2003.
- Lewis D. D., Yang T., Rose T., Li F., « RCV1 : A New Benchmark Collection for Text Categorization Research », *Journal of Machine Learning Research*, 2004.
- Lin W., Jin R., Hauptmann A., « Meta-classification of Multimedia Classifiers », *Proceedings of First International Workshop on Knowledge Discovery*, 2002.
- Naphade M., « On Supervision and Statistical Learning for Semantic Multimedia Analysis », *Journal of Visual Communication and Image Representation*, 2004.
- Over P., Kraaij W., Smeaton A., « TRECVID 2005 - An Introduction », *TRECVID 2005 - Text Retrieval Conference TRECVID Workshop*, 2005.
- Snoek C. G., Worring M., Geusebroek J.-M., Koelma D. C., Seinstra F. J., Smeulders A. W., « The Semantic Pathfinder for Generic News Video Indexing », *Proceedings of ICME*, 2006.
- Snoek C., Worring M., Smeulders A., « Early versus Late Fusion in Semantic Video Analysis », *In proceedings of ACM Multimedia*, 2005.
- Wolpert D. H., « Stacked Generalization », *Journal of Neural Networks*, 1990.