MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE
LA RECHERCHE SCIENTIFIQUE
UNIVERSITE ABDELHAMID IBN BADIS - MOSTAGANEM

Faculté des Sciences Exactes et de l'Informatique
Département de Mathématiques et d'Informatique
Filière : Informatique

MEMOIRE DE FIN D'ETUDES
Pour l'Obtention du Diplôme de Master en Informatique
Option : **Ingénierie des Systèmes d'Information**

THEME :

Descripteurs classiques vs. Descripteurs appris pour la détection de multi-concepts dans les images

Etudiant : « **KHOUSSA Mohamed El Bachir** »

Encadrant : « Dr. HAMADI Abdelkader »

Année Universitaire 2016/2017

Résumé

La détection de concepts visuels dans les images est une tâche très importante pour concevoir des systèmes de recherche sémantique d'images. Bien que cette problématique soit très difficile, les performances des approches proposées dans l'état de l'art s'améliorent. Or, indexer des documents par des concepts singuliers ne suffit pas pour répondre à des requêtes complexes des utilisateurs qui comportent plusieurs sémantiques. Il est donc important de penser à la problématique de détection de plusieurs concepts simultanément (multi-concepts) dans les images afin d'aboutir à des résultats de recherche plus satisfaisants. Cette tâche a été très peu abordée dans l'état de l'art. D'autre part, en plus des descripteurs classiques (de bas niveau) utilisés dans les systèmes d'indexation des images, d'autres types de caractéristiques de haut niveau ont émergé et ont donné des résultats intéressants. Ce genre de descripteurs sont extraits via une étape d'apprentissage, soit en utilisant l'apprentissage profond, soit en exploitant la détection de certaines sémantiques. Dans le cadre de ce travail, une étude comparative des deux types de descripteurs (Descripteurs de bas vs haut niveau) est réalisée dans le contexte de la détection des multi-concepts dans les images. Nous avons conduit une évaluation de nos contributions sur le corpus « Pascal VOC 2012 » pour la détection de paires et de triplets de concepts. Nous avons obtenu de très bons résultats rivalisant avec ceux de l'état de l'art.

Mots-clés : Indexation sémantique, multimédia, descripteurs de bas niveau, descripteurs de haut niveau, apprentissage supervisé, apprentissage profond, concepts multiples, bi-concepts, tri-concepts, Pascal VOC.

Dédicaces

Je dédie ce mémoire

À mes chers parents ma mère et mon père

À mes sœurs et mes frères

À mes adorables nièces et neveux

À toute ma famille

À mes amis

A tous ceux qui, par un mot, m’ont donné la force de continuer...

Remerciements

Je souhaite manifester mes sincères remerciements à Dieu, le tout puissant, pour ses faveurs et ses gratitudes de m'avoir donné le courage et la patience pour achever ce modeste travail.

Je tiens à présenter de tout mon cœur mes remerciements et mes reconnaissances à mon honorable encadreur Mr « HAMADI Abdelkader » pour son aide, ses conseils précieux, sa gentillesse, son encouragement, sa disponibilité et sa confiance qui m'a permis de ne jamais faiblir et de poursuivre toujours plus loin dans mon travail.

Je remercie naturellement ma famille pour son aide, sa générosité et son soutien moral qui ont été pour moi une source de courage et de confiance.

Je remercie aussi vivement les honorables membres du jury qui ont accepté d'évaluer ce travail.

Enfin, un grand remerciement à tous ceux qui, par un mot, m’ont donné la force de continuer à travailler afin d’atteindre mes objectifs.

Sommaire

Liste des Tableaux

Liste des Figures

Liste des Abréviations

- **SIFT :** Scale Invariant Feature Transform.
- **DCNN :** Deep Convolutional Neural Network.
- **Multi-concept :** un ensemble de concepts.
- **learnMulti :** Détection de multi-concepts par modèles de concepts multiples.
- **singleFus :** Détection de multi-concepts par fusion des détecteurs de concepts singuliers.

Introduction générale

L'explosion de la masse de données multimédia et les besoins des utilisateurs sont toujours en croissance, alors la nécessité d'un système de traitement et d'analyse automatique de cette masse de données n'est donc plus à démontrer.

Plusieurs systèmes d'indexation et de recherche de documents multimédia par le contenu ont vu le jour, la majorité d'entre eux concernent les images. Un système typique de recherche d'images par le contenu permet aux utilisateurs de formuler des requêtes en présentant un exemple du type de l'image recherchée. Le système identifie alors parmi la collection d'images celles qui correspondent le plus à l'image requête, et les affiche. La correspondance entre l'image requête et l'ensemble des images de la base de données se fait en comparant les caractéristiques de bas niveau des images qui sont des mesures mathématiques de la couleur, la texture et/ou de la forme …etc.

Malgré leur utilité, ce genre de systèmes ne répond pas à tous les besoins des utilisateurs. La manière la plus simple pour un utilisateur est de formuler ses attentes à travers des termes textuels. Pour ce faire, un système automatique doit pouvoir faire une correspondance entre du texte compréhensible par l'humain et un contenu multimédia. Autrement dit, il devient nécessaire de passer à une analyse sémantique. Cela constitue un des majeurs défis de la recherche d'information multimédia par le contenu. On distingue alors un autre genre de systèmes qui utilisent en plus des approches opérant sur un contenu de bas niveau, d'autres approches qui manipulent des sémantiques relatives aux documents. Ces dernières approches sont plus intéressantes mais elles présentent plusieurs difficultés.

Dans le cadre de la recherche d'information multimédia on assigne aux documents des concepts ou des termes sémantiques. Ce processus appelé « indexation sémantique » peut être réalisé de trois manières différentes [1] :

- Manuelle : par l'intervention d'un expert humain pour attribuer à chaque document multimédia un ou un ensemble de concepts/sémantiques qui lui sont associés. Avec l'explosion de la masse de données multimédia, cette méthode devient de plus en plus impossible à réaliser de façon entièrement manuelle ;
- Automatique : le processus est réalisé par une machine. Bien que cette méthode soit applicable sur une grande masse de données, la qualité d'indexation est très insuffisante pour avoir des recherches précises et efficaces ;
- Semi-automatique : c'est une solution intermédiaire entre les deux précédentes.

Parmi les difficultés et les défis de l'indexation sémantique on distingue :

- Le fossé sémantique : c'est l'un des problèmes majeurs lors de la construction d'un système de détection de concepts. Il est défini par la distance séparant les données brutes (les pixels) des images et les interprétations humaines de ces images ;
- Le fossé sensoriel : est défini comme le fossé existant entre le monde réel 3D et sa représentation en une image 2D. Lors de l'acquisition des images et vidéos, cette projection vers un espace 2D provoque une grande perte d'informations [1] ;
- Les classes déséquilibrées : il est très fréquent qu'une large majorité d'exemples d'apprentissage soient annotés par rapport à une seule des deux classes. La

performance des algorithmes standards d'apprentissage et de classification est sensiblement affectée par ce problème, parce qu'ils sont souvent basés sur l'optimisation de la précision ou du taux d'erreurs. Cette optimisation est biaisée par la classe majoritaire [1];

- D'autres problèmes tels que : le problème de la disponibilité de données, les difficultés de description (descripteurs pas efficaces ou peu efficaces pour décrire toutes les sémantiques).

Plusieurs efforts ont visé à réduire le fossé sémantique, et à traiter ou contourner certains autres problèmes rencontrés dans le domaine de l'indexation multimédia. Ces efforts ont été concrétisés par une amélioration significative de la performance des systèmes détection de concepts dans les images et/ou les vidéos.

En réalité, une requête d'un utilisateur est plus complexe qu'une représentation sémantique par un concept singulier. Il est donc important de penser à indexer les documents multimédia avec plus d'un seul concept singulier afin d'aider les systèmes de recherche à répondre à de telles requêtes complexes. En effet, un ensemble de concepts (multi-concept) peut représenter plusieurs sémantiques d'une simple combinaison de mots. Par exemple, la combinaison des concepts "neige", "montagne" et "personne(s) en train de se déplacer" pourrait être liée à une scène de "skieur" ou "une compétition de ski" [2].

Trop peu de travaux ont été réalisés pour étudier le problème de détection d'un ensemble de concepts (multi-concept) dans les images et vidéos. La majorité d'entre eux concerne la détection de pairs de concepts (bi-concept) [3]. Ces travaux montrent qu'un tel défi de détection de multi-concept est plus difficile que celui des concepts singuliers [2].

Les descripteurs de bas niveau connus dans l'état de l'art ne sont pas très efficaces pour représenter le contenu des images. En effet, bien que certains donnent de bons résultats pour détecter certains concepts, leurs résultats ne sont pas similairement bons pour tous les concepts. Ces dernières années, un autre type de descripteurs a vu le jour pour la détection de concepts : « Les descripteurs de haut niveau ». Ces descripteurs sont extraits d'un niveau supérieur à celui du signal (le niveau pixel). Généralement, ces descripteurs sont générés après une étape d'apprentissage. Cela les qualifie de sémantiques. Des travaux ont montré que ces descripteurs peuvent donner des performances comparables ou meilleures que celles des descripteurs de bas niveau de l'état de l'art dans la détection de concepts singuliers [4].

Dans notre travail nous nous intéressons à l'utilisation des descripteurs de haut niveau pour la détection de multi-concept et à la comparaison de leurs performances à celles des descripteurs de bas niveau. Nous avons cadré notre travail pour détecter des paires et des triplets de concepts (bi-concept et tri-concept) dans le contexte de la compagne d'évaluation Pascal VOC, en utilisant la collection Pascal VOC 2012[1].

Le reste du rapport sera divisé en cinq chapitres suivis d'une conclusion générale. Le chapitre 1 est dédié à la description de l'architecture d'un système de détection de concepts dans les documents multimédia et à l'étude détaillé de chaque partie le composant. Nous y présenterons également les approches de détection de multi-concepts dans les images. Dans

[1] Pascal Visual Object Classes (VOC) 2012 collection. http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

les chapitres 2 et 3 nous présenterons les descripteurs de bas niveau et de haut niveau, respectivement. Le chapitre 4 est dédié aux contributions, où nous présenterons nos différentes approches utilisées, ainsi qu'une présentation des données et des expérimentations faites, terminant par une discussion des résultats. Dans le chapitre 5 nous présenterons les différentes étapes de conception et d'implémentation de notre système de détection de concept réalisé.

Chapitre 1

Système de détection de concepts

1.1. Introduction

Dans ce chapitre nous allons décrire la structure de base d'un système de détection de sémantiques (concepts) dans les images, en expliquant les différentes étapes de sa construction, et les approches utilisées dans l'état de l'art pour améliorer la performance de ce type de systèmes. Le système décrit reste valable pour le cas des vidéos.

1.2. Architecture d'un système de détection de concepts

L'indexation sémantique de documents multimédia est généralement réalisée par détection de concepts visuels via des approches d'apprentissage automatique supervisée. Un classificateur est entrainé sur un ensemble de données annotées manuellement par rapport un concept cible. Une classification binaire est réalisée (Deux classes : la classe positive pour les images qui contiennent le concept cible, et la classe négative des échantillons qui ne contiennent pas le concept étudié).

Pour chaque échantillon de l'ensemble d'entrainement, un ou plusieurs descripteurs de bas niveau sont extraits. On peut réaliser une fusion de plusieurs descripteurs d'une image en les agrégeant pour concevoir un seul descripteur (fusion précoce). Les descripteurs considérés passent par une chaine de prétraitements qui inclut l'opération de normalisation des valeurs du descripteur et la réduction de dimensionnalité de ce dernier. Chaque exemple est représenté par le couple (descripteur prétraité, annotation) où l'annotation $\in \{0, 1\}$ est son annotation manuelle par rapport au concept cible. L'annotation prend deux valeurs possible : 1) la valeur 1 signifie que le concept cible est présent dans l'échantillon ; 2) la valeur 0 signifie que l'échantillon ne contient pas ce concept. A l'issue de la phase d'apprentissage, un modèle est généré. Ce modèle sera utilisé après, pour prédire si des documents non encore vus contiennent le concept étudié. Un score est calculé pour refléter la probabilité que l'échantillon contient le concept. On peut réaliser une fusion tardive de plusieurs descripteurs utilisés séparément en agrégeant leurs scores respectifs. Les scores finaux de détection sont utilisés pour ordonner les échantillons dans le but d'évaluer la performance des systèmes. Pour améliorer la performance on peut changer cet ordre en appliquant une méthode de ré-ordonnancement qui utilise généralement des informations contextuelles ou externes.

La Figure 1.1 présente l'architecture d'un système de détection de concepts dans les documents multimédia [1]. Nous allons décrire par la suite chaque étape de ce système.

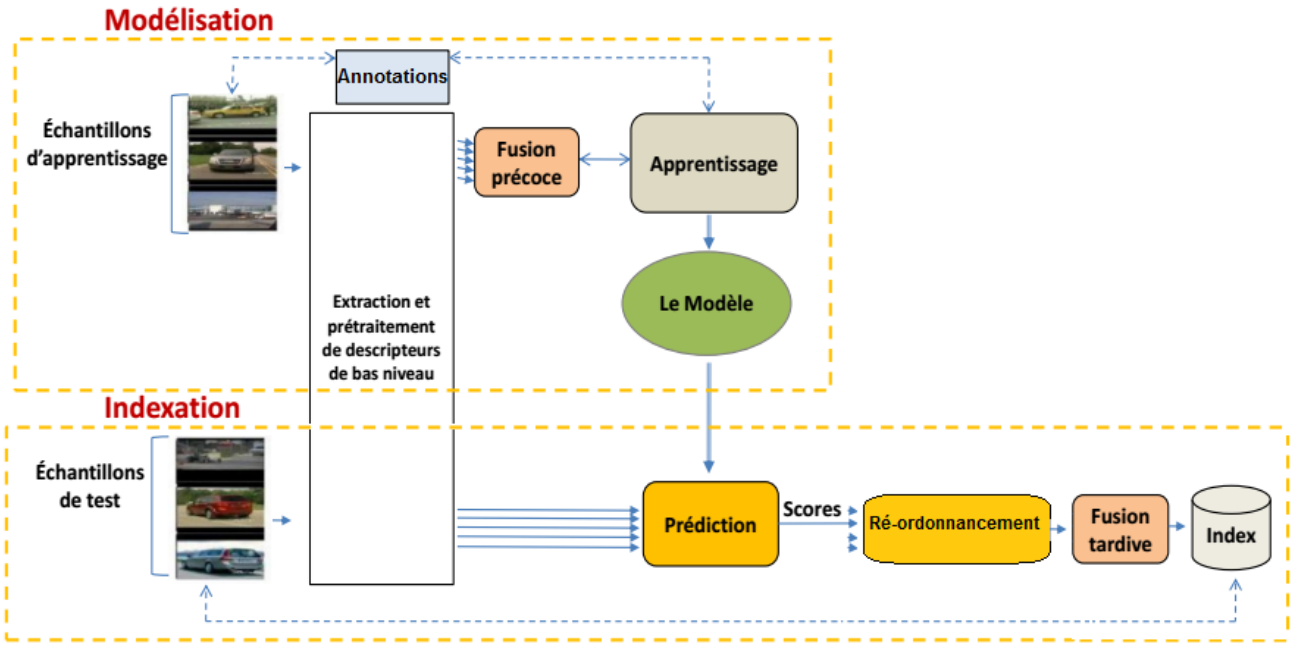

Figure 1.1 Architecture d'un système de détection de concepts dans les documents multimédia [1]

1.3. Description des images

Dans un système de détection de concepts dans les images, il est nécessaire de passer par une opération d'extraction de caractéristiques visuelles des images, qui permet d'obtenir une représentation plus facilement manipulable. Les pixels constituent une première représentation des images mais très peu robustes aux variations de luminosité, échelle, rotation …etc.

Il n'existe pas une définition précise d'un descripteur, tout dépend du problème et du type d'application. Les descripteurs sont en général utilisés en entrée des algorithmes de la vision par ordinateur. Cela rend l'efficacité de ces algorithmes relative à la qualité des techniques d'extraction des descripteurs utilisés. Un bon descripteur doit permettre de reconnaitre le contenu même en cas de certaines variations : changement d'illumination, variation de l'échelle, translation et rotation, changement de points de vue …etc. [1].

1.3.1. Catégorisation des descripteurs

On peut catégoriser les descripteurs des images selon plusieurs critères.

Descripteurs globaux / locaux

- Descripteurs globaux : extraits à partir de l'image entière ;
- Descripteurs locaux : extraits sous forme d'un ensemble de caractérisations pour chaque primitive de l'image, c'est à dire par zones d'intérêt (points, blocs, régions …etc.). Par exemple, la couleur pour chaque région de l'image.

Descripteurs bas niveau / haut niveau

- Descripteurs de bas niveau : extraits à partir du signal (pixels). (Voir Chapitre 2) ;
- Descripteurs de haut niveau : ces descripteurs modélisent et manipulent des informations de type sémantique. Ils sont extraits via une étape d'apprentissage. Ce sont des descripteurs sémantiques et/ou appris. (Voir Chapitre 3) ;

Il y a aussi d'autres catégorisations possibles. Par exemple, selon le type d'information manipulée : la couleur, la texture, la forme ou les points d'intérêt (Voir Chapitre 2).

1.3.2. Optimisation des descripteurs

Avant d'utiliser un descripteur dans un système d'apprentissage ou de classification, il est recommandé de le soumettre à une chaine de prétraitements. Cette phase est appelée : "optimisation de descripteurs", elle passe par deux étapes importantes :

- La normalisation des descripteurs qui vise d'une part à uniformiser la distribution des valeurs du vecteur caractéristique, de façon à ce qu'il n'y ait pas un grand écart entre les différentes valeurs. Cela s'avère très utile pour réduire l'influence des grandes valeurs qui dominent les petites valeurs. Et d'autre part, à étaler l'ensemble de valeurs, de manière à ce qu'elles couvrent le maximum possible d'un intervalle donné.
- La réduction de dimensionnalité quant à elle, et comme son nom l'indique, est une méthode permettant de réduire le nombre de composantes formant le descripteur, par la projection des données dans un autre espace de dimension inférieure, sans écarter de l'information significative, ou pour être plus précis, en gardant le maximum possible d'informations, car cette projection va causer une perte d'informations, dépendamment du nombre et du choix des dimensions éliminées.

Un descripteur optimisé a tendance à être plus efficace dans une approche de classification. Il est aussi possible d'appliquer une seule des deux méthodes avant d'utiliser le descripteur (i.e. ou bien normaliser le descripteur ou bien réduire sa dimensionnalité) [1].

1.4. Modélisation et Prédiction (Indexation)

Généralement, dans un système de détection de concepts dans les documents multimédias, la classification est réalisée de façon binaire (Voir Section 1.2). Le processus se déroule en deux phases. La première est une étape de modélisation qui se fait par apprentissage supervisé sur un ensemble d'échantillons annotés par le concept cible. Cette phase génère un modèle. Ensuite, dans une seconde étape, le système donne des prédictions sur un ensemble d'échantillons de test à l'aide du modèle généré dans la phase précédente. À la fin de ce processus, on obtient un ensemble d'échantillons indexés automatiquement par le concept cible.

Afin de pouvoir faire une classification binaire dans un système de détection de concepts dans les documents multimédias, plusieurs méthodes peuvent être appliquées (Réseaux de neurones, SVM ...etc.). Parmi ces différentes méthodes, SVM semble être une des plus performantes dans l'état de l'art [1], mais ses résultats ne sont pas encore satisfaisants. En

plus de ça, les corpus de données sont déséquilibrés où la classe négative est dominante. Pour cela, une solution consiste à un sous-échantillonnage du corpus avec une approche de Bagging. Une telle approche utilisant SVM (multi-SVM) a été proposée dans [5] et a donné de très bons résultats pour la détection de concepts singuliers dans les plans de vidéos.

- Multi-SVM

Dans [5], les auteurs proposent un schéma de Bagging basé sur les SVM avec une sélection biaisée des échantillons positifs et négatifs pour traiter le problème des classes déséquilibrées dans le cadre de la détection de concepts dans les documents multimédia. Cette méthode consiste à combiner *m* classificateurs via une stratégie de "Bagging" où chacun d'entre eux utilise tous les échantillons d'apprentissage de la classe dominée (typiquement, la classe positive) et un ensemble d'échantillons de la classe dominante (typiquement, la classe négative) est tiré aléatoirement avec remise (boostrap), avec :

$$m = (f_{neg} \times N_{neg})/(f_{pos} \times N_{pos})$$

Où N_{pos} est le nombre d'échantillons positifs, N_{neg} est le nombre d'échantillons négatifs, f_{neg} et f_{pos} sont des paramètres (entiers positifs non nuls) relatifs aux classes positive et négative, respectivement. f_{pos} gère la proportion des échantillons de la classe dominante qu'on veut utiliser, par rapport au nombre d'exemples de la classe dominée (ex. Deux fois plus d'exemples négatifs que positifs). f_{neg} quant à lui, permet de contrôler, à l'aide de f_{pos}, le nombre de classificateurs souhaité. L'ensemble des échantillons d'apprentissage est divisé en *m* sous-ensembles, où chaque sous-ensemble contient tous les exemples positifs contenus dans cet ensemble et ($f_{pos} \times N_{pos}$) exemples négatifs sont tirés aléatoirement avec remise. Ensuite chacun des *m* classificateurs est entrainé sur un sous-ensemble différent. On remarque que la contrainte $f_{neg} \times N_{neg} \geq f_{pos} \times N_{pos}$ doit être vérifiée. Finalement, les scores des *m* classificateurs sont fusionnés en utilisant n'importe quelle fonction possible, typiquement une moyenne arithmétique. Plus la valeur de *m* est grande, meilleure est la performance finale. Il a été montré qu'utiliser SVM comme classificateur de base donne de meilleurs résultats dans le domaine de la détection de concepts dans les images [1].

1.5. Détection de multi-concepts

Beaucoup d'efforts et de travaux ont été fait pour améliorer la performance des systèmes d'indexation des images/vidéos, mais la majorité d'entre eux concerne la détection de concepts singuliers. Cependant, avec l'évolution des systèmes de recherche d'information, les besoins des utilisateurs sont en augmentation et deviennent plus complexes en termes d'abstraction et en nombre de mots composants les requêtes. Par conséquent, si les images et les vidéos sont indexées avec des groupes de concepts (des multi-concepts), les performances des systèmes de recherche vont s'améliorer. En outre, trouver une description visuelle de l'occurrence de multi-concept devient un défi très difficile. Cette remarque est confirmée par les résultats médiocres des participants à la tache de détection de paires de concepts proposée

par TRECVid[2] en 2012 et 2013, par rapport aux résultats obtenus pour la détection de concepts singuliers [2].

On distingue deux approches possibles pour détecter simultanément un groupe de concepts [1] :

- Modèles de concepts multiples : cette approche consiste à générer un modèle spécifique pour chaque groupe de concepts. Autrement dit, elle considère le groupe de concepts comme un nouveau concept et elle génère un modèle spécifique pour chaque multi-concept. Pour instancier ce modèle, on se base uniquement sur les données relatives aux concepts singuliers. Etant donné un ensemble d'échantillons annotés par un ensemble de concepts singuliers, nous générons les annotations des mêmes échantillons par un multi-concept en réalisant une intersection des annotations par les concepts singuliers qui le composent. Pour faire cela on utilise une fonction qui prend en entrée un échantillon et un multi-concept et renvoie l'annotation de cet échantillon par rapport au multi-concept. Cette fonction renvoie 1 qui désigne « l'occurrence du multi-concept dans l'échantillon » si tous les concepts singuliers composant le multi-concept sont annotés par 1 dans cet échantillon, et elle renvoie -1 qui désigne « pas d'occurrence du multi-concept dans l'échantillon » s'il existe un des concepts singuliers qui composent le multi-concept annoté par -1, sinon elle renvoie 0 qui désigne « pas de décision ». À la fin, on obtient un ensemble annoté par le multi-concept à détecter. À base de ces annotations, on peut dérouler le processus du système de détection de concepts comme pour la détection de concepts singuliers. La Figure 1.2 montre un exemple de génération des annotations pour un ensemble d'images, par une paire de concept (bi-concept) composé des concepts singuliers concept-1 et concept-2, pour lesquels on détient les annotations pour l'ensemble des images considéré.

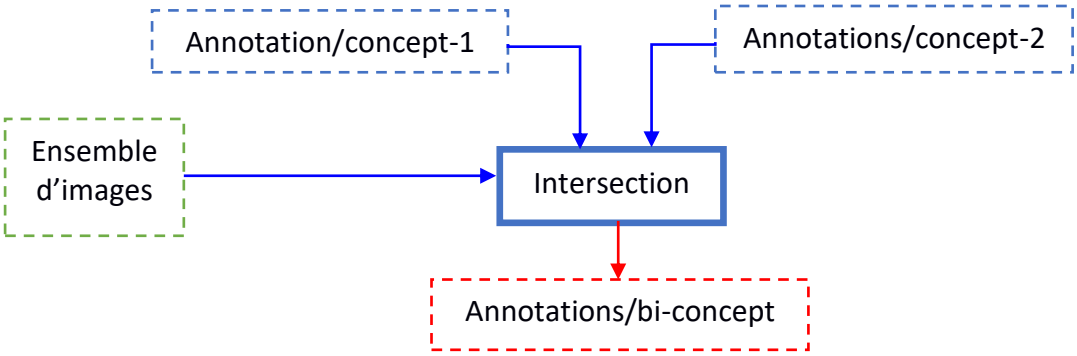

Figure 1.2 Génération des annotations d'un bi-concept

[2] http://www-nlpir.nist.gov/projects/trecvid/

- Fusion de détecteurs de concepts individuels : cette approche consiste à détecter l'ensemble des concepts formant le groupe séparément, c'est-à-dire, dérouler le processus du système de détection de concepts pour chaque concept singulier qui appartient au multi-concept cible. Ensuite, combiner les résultats de leurs détections respectives, en utilisant une méthode de fusion afin d'obtenir le résultat (score) de détection du multi-concept. Plusieurs fonctions de fusion de scores de détection des concepts singuliers peuvent être utilisées. On peut citer :
  - Une fusion linéaire qui renvoie une moyenne arithmétique des scores de détection des concepts singuliers ;
  - Une méthode basée sur la notion de probabilité. Cette méthode considère que les scores sont des probabilités et que ces probabilités sont obtenues par des détecteurs conditionnellement indépendants. Elle renvoie la racine nième (n étant le nombre de concepts singuliers composant le multi-concept) du produit des scores de détection des concepts singuliers ;
  - Une approche inspirée de l'approche booléenne étendue, qui considère un multi-concept comme la conjonction des concepts qui le composent.

  La Figure 1.3 montre un exemple de génération des scores de détection d'une paire de concepts (bi-concept) composée des concepts singuliers concept-1 et concept-2. Les scores de détection de ces derniers pour l'ensemble des images considéré, sont obtenus en déroulant le processus de détection de concepts pour chacun d'entre eux séparément.

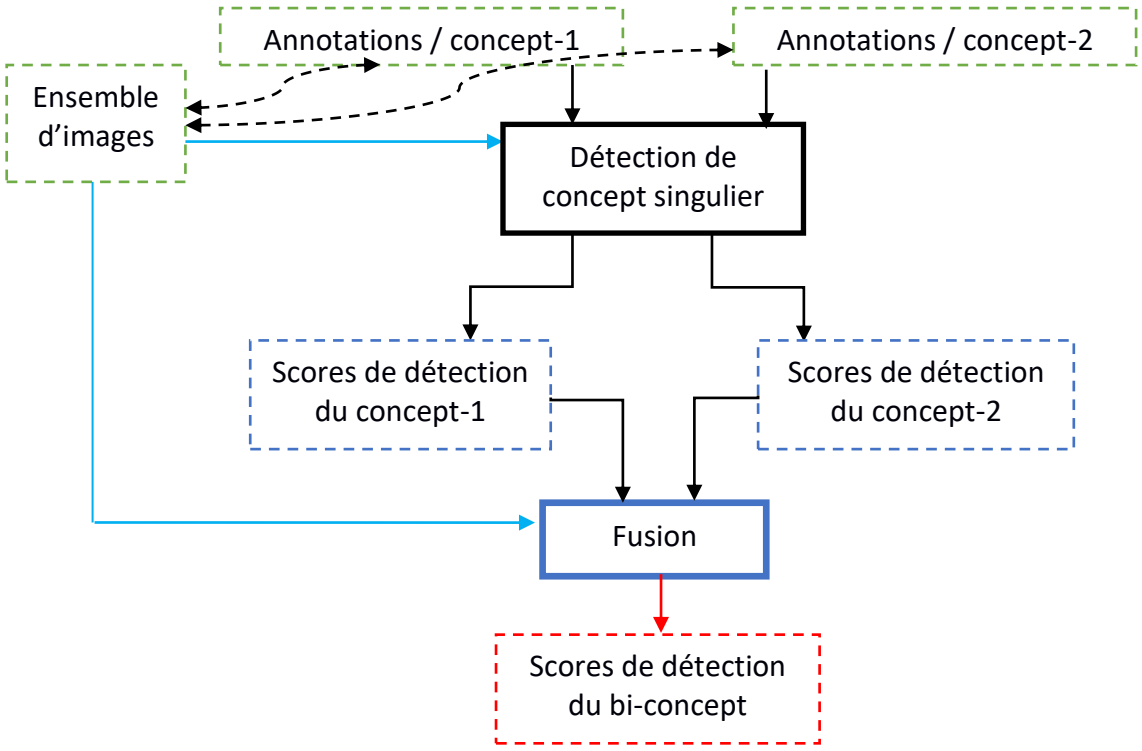

Figure 1.3 Génération des scores de détection d'un bi-concept

1.6. Fusion

Généralement un seul descripteur ne suffit pas pour avoir des performances satisfaisantes dans le cadre de la détection de concepts. En effet, trouver un bon descripteur pour décrire chaque classe ou un concept reste un défi ouvert. Cette remarque reste vraie pour les données multimédia, pour lesquelles il y a plusieurs sources d'information. Par exemple, on peut décrire une image en se basant sur la couleur, la texture, la forme, ou en extrayant des points d'intérêts. Il est nécessaire de prendre en compte toute information utile pour une bonne description des données. Pour ce faire, des chercheurs ont étudié la combinaison des informations de diverses sources dans le but d'améliorer les performances de leurs systèmes d'indexation et/ou de classification. Cette combinaison est appelée "fusion". La fusion peut être appliquée sur deux niveaux différents (bas niveau / haut niveau) [1] .

1.6.1. Fusion de bas niveau « fusion précoce »

La fusion précoce (*early fusion* en anglais) consiste à combiner un ou plusieurs descripteurs uni-modal(aux) pour générer une nouvelle représentation regroupant des informations issues de différentes modalités. Dans ce cas, les informations fusionnées sont dites brutes, c'est-à-dire proches du "signal". Dans le cadre de l'indexation et/ou classification, on parle de fusion de descripteurs de bas niveaux. Cela revient en pratique à concaténer plusieurs descripteurs de bas niveau extraits de différents médias (ex. Audio, visuel), ou plusieurs descripteurs de différents types (ex. Couleur, texture, …etc.) pour générer un nouveau vecteur de dimension plus grande. Ce nouveau descripteur est ensuite utilisé dans une approche d'apprentissage.

1.6.2. Fusion de haut niveau « fusion tardive »

Contrairement à la fusion précoce qui combine des informations de bas niveau, la fusion tardive (*late fusion* en anglais) fusionne des informations sémantiques, qui sont souvent des scores de classification/prédiction ou des probabilités renvoyés par des apprenants entrainés sur différents descripteurs. Les scores de prédiction sont fusionnés en utilisant une simple fonction (moyenne, min, max …etc.) ou via un nouvel apprenant (classificateur). Notons ici que ce sont les résultats des prédictions qui sont fusionnés : ces derniers ne sont pas nécessairement liés à des différents descripteurs de bas niveau. En effet, on peut lancer plusieurs classificateurs différents en utilisant un même descripteur et fusionner les résultats obtenus.

1.7. Ré-ordonnancement

L'indexation sémantique est généralement réalisée par un apprentissage supervisé, où le système est entrainé sur les échantillons positifs et négatifs par rapport à un concept cible (l'ensemble de développement) pour produire un modèle qui est alors utilisé pour la production des scores reflétant les probabilités que de nouveaux échantillons (l'ensemble de test) contiennent ce concept. Ces scores sont souvent calculés de façon homogène à une probabilité. La recherche peut alors être effectuée en classant les échantillons en fonction du score de probabilité, de façon à mettre dans le haut de la liste les échantillons les plus

susceptibles de contenir le concept cible. Il est souvent possible d'améliorer le rendement d'indexation ou de détection en modifiant les scores de probabilités de l'ensemble des échantillons, et ce en utilisant les scores initiaux ainsi que d'autres sources d'information. Les nouveaux scores engendrent un nouveau classement des échantillons. Le ré-ordonnancement est connu dans l'état de l'art aussi sous d'autres noms : reclassement, re-scoring et re-raking. Récemment, plusieurs travaux se sont focalisés sur les méthodes de ré-ordonnancement (utilisation des résultats de différents modèles de recherche, utilisation de nouveaux modèles de classification pour faire le ré-ordonnancement …etc.) [6].

1.8. Utilisation du contexte pour la détection de concepts

Il existe plusieurs définitions du contexte. Dans [1], le contexte est défini comme suit : *«le contexte est toute information additionnelle qu'un système d'indexation de base peut s'en passer, qui est pertinente et peut aider à améliorer la qualité de l'indexation»*. Ceci dit, si la présence ou l'absence d'une information est pertinent pour un système par rapport à son objectif final, alors cette information est considérée comme un contexte. L'information contextuelle peut être intégrée à plusieurs niveaux possibles, dépendant du domaine abordé. Il y'a plusieurs possibilités différentes d'utiliser le contexte dans un système d'indexation de documents multimédia, et il peut se faire dans des différentes étapes du système (prétraitement de données, extraction et optimisation des descripteurs (avant l'apprentissage), pendant l'apprentissage …etc.). Plusieurs travaux ont montré que le contexte pourrait améliorer significativement la performance des systèmes de détection de concepts dans les images et les vidéos [1].

1.9. Évaluation d'un système de détection de concepts

Les résultats d'un système de détection de concepts nécessitent une évaluation de leurs qualités. On a aussi besoin de comparer les performances d'un système par rapport à certaines autres. Il existe plusieurs mesures d'évaluation dans l'état de l'art. D'autre part, des compagnes d'évaluation dans le domaine de l'indexation multimédia ont vu le jour.

1.9.1. Mesures d'évaluation

Les mesures d'évaluation les plus populaires, pour la comparaison des systèmes de recherche d'information sont la précision et le rappel. Ces métriques sont largement utilisées pour l'évaluation de l'efficacité des approches d'annotation automatique, dans la communauté de la recherche d'information. Dans cette dernière, la précision d'une requête est définie par le ratio du nombre des documents pertinents retournés par le système et le total du nombre de documents retournés. Le rappel quant à lui, est défini par le ratio des documents pertinents retournés par le système et le nombre total des documents pertinents dans la base de données.

- $précision = \frac{|\{documents\ pertinents\} \cap \{documents\ retournés\}|}{\{documents\ retournés\}}$

- $rappel = \frac{|\{documents\ pertinents\} \cap \{documents\ retournés\}|}{\{documents\ pertinents\}}$

Il existe une autre métrique appelée « précision moyenne » qui a l'avantage de considérer l'ordre dans lequel les documents sont retournés

- Précision moyenne (Average Precision ou AP) : mesure la précision moyenne non interpolée.

L'AP a l'avantage de résumer la courbe "rappel-précision" en une seule valeur. Elle est largement utilisée comme la mesure officielle de plusieurs compagnes de recherche d'images et de vidéos, comme TRECVid et Pascal VOC. L'AP est définie par la formule suivante.

$$AP = \frac{1}{R}\sum_{j=1}^{S}\frac{P_j}{j} \times I_j$$

Où R est le nombre de documents pertinents dans un corpus contenant S documents. À chaque indice j, P_j est la précision après j documents qui sont retournés, et I_j est égal à 1 si le document à l'indice j est pertinent, et 0 sinon.

1.9.2. Compagnes d'évaluation

Plusieurs compagnes d'évaluation ont vu le jour dans le domaine de l'indexation multimédia. Ces dernières organisent des compétitions internationales inter-laboratoires et/ou équipes de recherche, en fournissant des corpus de données et éventuellement des outils de travail. Les équipes participantes fournissent leurs résultats d'indexation sur un corpus de test (pour lequel les on ne connait pas la vérité-terrain) et les organisateurs se chargent d'évaluer les performances des résultats des participants.

- Pascal-VOC : PASCAL Visual Object Classes (VOC) est un défi qui est devenu une référence dans le domaine de la détection et la reconnaissance des catégories d'objets visuels. Les données des différentes années sont disponibles en ligne sur le site officiel du défi[3].
- TRECVid : Depuis 2001, la compagne d'évaluation TREC VIDEO offre à ses participants les moyens pour expérimenter les différentes approches de détection de concepts dans les vidéos[4].
- ImageCLEF : ImageCLEF a été lancée en 2003 dans le cadre du Forum d'évaluation inter-langues (CLEF : Cross Language Evaluation Forum). Elle vise à fournir un forum d'évaluation pour l'annotation inter-langues et la recherche des images[5].

Dans notre projet nous allons évaluer nos approches utilisées sur le corpus Pascal-VOC 2012.

[3] http://host.robots.ox.ac.uk/pascal/VOC/
[4] http://www-nlpir.nist.gov/projects/trecvid/
[5] http://www.imageclef.org/

1.10. Conclusion

A travers ce chapitre nous avons pu définir un système de détection de concepts. Nous avons détaillé les différentes étapes de ce système. Par la suite nous avons présenté les différentes mesures utilisées pour évaluer et valider un tel système.

Le chapitre suivant est dédié à la présentation des descripteurs de bas niveau pouvant être extraits d'une image.

Chapitre 2

Descripteurs d’images de bas niveau

2.1. Introduction

Un système de détection de concepts dans les images intègre une partie d'extraction de descripteurs. De façon générale, les documents multimédia dont on veut extraire l'information sont stockés au sein de bases de données dans des formats permettant leur lecture (ou affichage), par un logiciel de diffusion. Ces formats ne contiennent pas d'information sur le sens du document. Afin de traiter de la sémantique, l'utilisation des descripteurs, c'est-à-dire de données qui décrivent le contenu des documents multimédia, devient incontournable. Les descripteurs des images sont extrêmement variés, et on peut les classer suivant divers critères. Dans la suite de ce chapitre nous allons décrire les différents descripteurs d'images de bas niveau.

2.2. Descripteurs de couleur

Les descripteurs de couleurs sont les plus utilisés dans l'état de l'art. La couleur forme une partie significative de la vision humaine. Il y a plusieurs descripteurs de couleur parmi eux on distingue :

- L'histogramme : c'est un descripteur global très courant dans l'état de l'art, qui représente la répartition de niveaux de gris (ou de couleurs) dans une image. L'histogramme s'avère le descripteur le plus simple à calculer, son calcul consiste à compter le nombre d'occurrences des différentes valeurs possibles d'intensité des pixels dans l'image. On peut distinguer plusieurs catégories d'histogrammes qu'on peut classer, par exemple, selon l'espace de couleur considéré lors du calcul : "histogramme RGB", "histogramme HSV", "histogramme Opponent", associés respectivement aux espaces de couleurs : "RGB", "HSV", "Opponent color space" [1]. L'histogramme (s'il est normalisé) est invariant à des modifications globales de l'image, telles que (la translation, la rotation et le changement d'échelle).

- Le corrélogramme : il a été proposé pour qualifier non seulement la distribution de couleurs des pixels, mais aussi la corrélation spatiale entre les pairs de couleurs. Il recherche des motifs dans un voisinage donné. Il est assimilable à une matrice ($n \times n \times r$) où n est le nombre de couleurs utilisées et r la distance maximale du voisinage considéré. Dans cette matrice, la valeur en (i, j, k) désigne la probabilité de trouver un pixel de couleur i *à* une distance k d'un pixel de couleur j. La représentation se fait le plus souvent par un vecteur résultant de la concaténation des lignes de la matrice [1].

- Autres descripteurs de couleur : il y a plusieurs autres descripteurs de couleurs, on peut en citer : l'histogramme de couleur-structure CS, les moments statistiques, le vecteur de cohérence de couleurs, le descripteur de couleurs dominantes, la distribution spatiale de couleur, la cohérence spatiale …etc.

2.3. Descripteurs de la texture

La texture représente également un descripteur de bas niveau efficace utilisé dans le cadre de l'indexation et la recherche par le contenu. Plusieurs techniques ont été développées pour mesurer la similarité de textures. La majorité des techniques comparent les valeurs de ce qui est connu par les statistiques du second ordre, calculées à partir des images requêtes. Ces méthodes calculent les mesures de textures d'images comme étant le degré de contraste, la grossièreté, la directivité et la régularité ; ou de la périodicité, la directivité et l'aspect aléatoire [1]. Et parmi les méthodes d'analyse de texture pour la recherche d'images on distingue :

- Le filtre de Gabor : le filtre de Gabor (ou ondelettes de Gabor) est largement adopté pour extraire les caractéristiques de textures à partir des images pour la recherche d'images. L'utilisation d'un banc de filtres de Gabor permet d'extraire de l'image considérée des informations pertinentes, à la fois en espace et en fréquence, relatives à la texture. En effet, plusieurs recherches conduites montrent que les fonctions de Gabor simulent de manière convenable le système visuel humain en reconnaissance des textures ; le système visuel étant considéré comme un ensemble de canaux de filtrage dans le domaine fréquentiel. La convolution de l'image par les filtres de Gabor peut se faire dans le domaine spatial ou fréquentiel [1].

- Statistiques du deuxième ordre : il y a des méthodes d'analyse de texture qui utilisent les statistiques du second ordre tel que (Transformée de Fourier, covariance …etc.) [1].

2.4. Descripteurs de formes

Les descripteurs de formes permettent, comme leur nom l'indique, de présenter une information pertinente sur le contenu de l'image et précisément sur la forme. Il existe différents types de descripteurs de formes qui se différencient par leur simplicité/complexité, comme les CSS (Curvative Scale Space descriptors) et les filtres de convolution [1]. Parmi les approches utilisées dans les descripteurs de formes on distingue :

- L'approche contour : elle décrit une région au moyen des pixels situés sur son contour. Elle fait classiquement référence aux descripteurs de Fourier et porte sur une caractérisation des frontières de la forme.
- L'approche région : elle considère une région par rapport aux caractéristiques des pixels de cette région, elle fait classiquement référence aux moments invariants et est utilisé pour caractériser l'intégralité de la forme d'une région.

Parmi les descripteurs de formes on peut citer par exemple : les moments géométriques, les moments orthogonaux, les descripteurs de Fourier…etc.

2.5. Descripteurs de points d'intérêt

L'extraction des descripteurs visuels sur l'image entière (descripteurs globaux) permet de réduire le nombre de calculs nécessaires, la taille de la base de données ainsi que le coût de recherche des images les plus similaires. Cependant, l'approche globale ne permet pas une recherche efficace d'objets (au sens large) dans l'image. À l'inverse, les descripteurs extraits d'une partie de l'image (descripteurs locaux) sont efficaces, mais coûteux. Les descripteurs locaux peuvent être des régions de l'image obtenues soit par segmentation de l'image entière (par recherche des régions d'intérêt) ou par recherche des points d'intérêt. Une manière de déterminer les points d'intérêt est de prendre en compte les zones où le signal change. Par exemple, les points d'intérêt peuvent être les coins, les jonctions en T ou les points de fortes variations de texture. Et parmi les descripteurs qui utilisent les points d'intérêt on distingue les SIFT et leur représentation en sacs de mots visuels.

- SIFT : Lowe [7] propose des descripteurs appelés SIFT (Scale Invariant Feature Transform), qui sont particulièrement utiles grâce à leur grande distinction dans le cadre de la reconnaissance. Ils sont obtenus en construisant un vecteur de grande dimension représentant les gradients dans une région locale de l'image. Les points d'intérêt de l'image sont calculés en utilisant un détecteur (le détecteur de Harris par exemple) à partir desquels les descripteurs SIFT seront calculés. En considérant un point d'intérêt P, le voisinage de P est décomposé en 16 blocs de 4x4 pixels (Voir Figure 2.1). Dans chaque bloc un histogramme d'orientation de gradients est formé, en discrétisant l'orientation en 8 bins (correspondant aux différentes orientations possibles: {0˚, 45˚, 90˚, 135˚, 180˚, 225˚, 270˚, 315˚}). Le vecteur représentatif du descripteur aura dans ce cas 128 composantes (4x4x8=128). Les descripteurs SIFT se montrent invariant à l'échelle et à la rotation et robuste au bruit et au changement de l'illumination. Les travaux réalisés dans le cadre de l'indexation de documents multimédia ont montré que ces descripteurs donnent pratiquement de meilleurs résultats, malgré leur grande dimension.

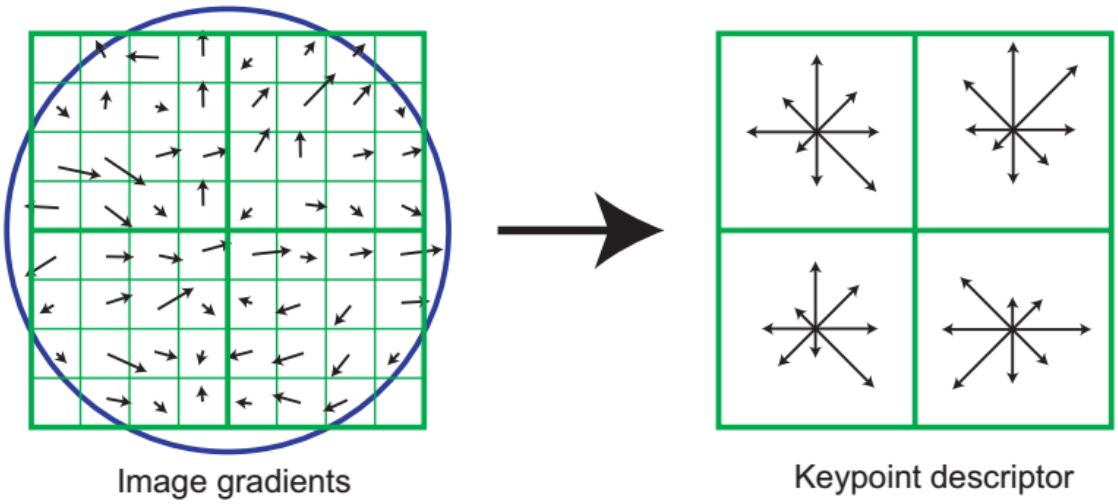

Figure 2.1 Principe de calcul des descripteurs SIFT [7]

- Sacs de mots visuels (utilisant SIFT): La représentation en sac de mots est une méthode qui consiste à représenter un document par l'ensemble de mots qui le constituent et qui appartiennent à un dictionnaire prédéfini de mots, connu aussi sous les noms de « codebook » ou « vocabulaire ». C'est une technique très réputée et largement utilisée dans le domaine de la recherche textuelle. Dans le cas des images et des vidéos on parle de sac de mots visuels. Le vocabulaire est construit en utilisant un ensemble d'apprentissage *Dev*. Des caractéristiques visuelles sont extraites des éléments de l'ensemble *Dev*, il en résulte un très grand nombre de points d'intérêts. Ce grand nombre est important pour la robustesse de la classification, mais évoque un défi de représentation à cause de la grande dimensionnalité (SIFT est d'une dimension égale à 128). Un regroupement (Clustering) est appliqué sur l'ensemble des points d'intérêts résultants, comme par exemple, la méthode "K-means" qui est très utilisée dans ce contexte. On obtient donc un ensemble de groupes (clusters), le centre de chacun d'entre eux représente un mot visuel. L'ensemble de ces mots visuels constitue le "vocabulaire" ou le "codebook". La représentation finale d'une image est l'histogramme, c'est à dire, les fréquences des mots visuels dans cette image. Cela est réalisé en faisant une correspondance entre les éléments du dictionnaire dont on dispose et chaque caractéristique extraite de l'image requête pour sélectionner le mot visuel le plus similaire à chaque caractéristique, et compter ensuite les fréquences [8].

- Autres descripteurs de points d'intérêts : il existe plusieurs autres descripteurs de points d'intérêt, on peut en citer : Opponent SIFT, OCLBP - Orthogonal Combination of Local Binary Pattern, ORB, SURF…etc [4] [9] [10].

2.6. Conclusion

A travers ce chapitre nous avons pu décrire les différents descripteurs d'images de bas niveau qui peuvent être utilisés dans un système de détection de concepts. Chacun de ces descripteurs pourrait subir une chaine d'optimisation pour une exploitation optimale comme indiquer dans (Section 1.3.2). Nous avons détaillé ces descripteurs selon le type d'information manipulée en donnant des exemples de chaque type.

Le chapitre suivant est dédié à la présentation des descripteurs de haut niveau pouvant être extraits d'une image.

Chapitre 3

Descripteurs d’images de haut niveau

3.1. Introduction

Les descripteurs d'images de bas niveau présentent des inconvénients : ils ne sont pas toujours efficaces. Un autre type de descripteurs d'images a vu le jour et les travaux de l'état de l'art qui l'ont utilisé ont abouti à de bons résultats. On parle de descripteurs de haut niveau dont on distingue deux approches pour les générer. La première méthode donne des descripteurs sémantiques basés sur la détection d'un ensemble de concepts pour détecter un concept cible. La deuxième approche quant à elle, donne des descripteurs appris basés sur une phase d'apprentissage profond sur les pixels de l'image. Dans la suite de ce chapitre nous allons décrire ces différents descripteurs d'images de haut niveau.

3.2. Descripteurs sémantiques

Les descripteurs sémantiques expriment des propriétés qui ont une signification forte. Ce sont des concepts qui peuvent être décrits en terme d'autres concepts. Leur construction nécessite le calcul de modèles d'apprentissage utilisant des descripteurs de bas niveau.

Dans un système de détection de concepts, le traitement utilisant des descripteurs de bas niveau est fait d'une manière complètement indépendante pour tous les échantillons et pour tous les concepts cibles. Une telle approche ne prend pas en compte les relations sémantiques et/ou statistiques entre les concepts cibles. Comme alternative à cette approche, on a une autre approche qui utilise les résultats de détection de plusieurs concepts pour construire un descripteur de détection d'un concept cible. Dans cette dernière, la phase d'apprentissage est faite sur plusieurs concepts différents du concept cible à détecter, ensuite on prend les résultats de détection de ces concepts pour construire un descripteur qui sera utilisé dans la détection du concept cible. Ce qui donne à cette approche la prise en considération des relations sémantiques et/ou statistiques entre les concepts. Cette approche de génération de descripteurs sémantiques est illustrée dans la Figure 3.1.

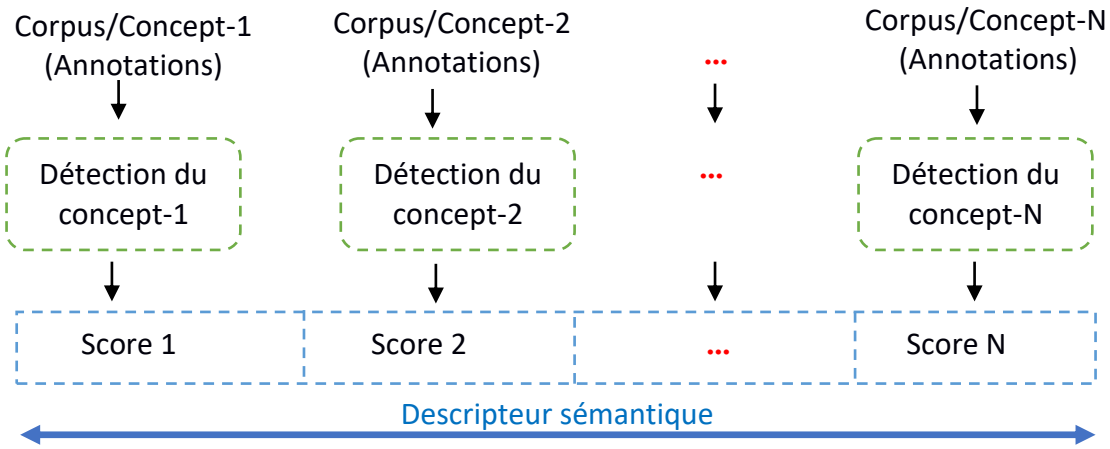

Figure 3.1 Phases de construction d'un descripteur sémantique

Dans ce qui suit on va citer quelques descripteurs sémantiques :

- Descripteur sémantique utilisant l'approche de rétroaction conceptuelle [11]: Pour détecter un concept cible on construit un descripteur conceptuel qui est une version normalisée du vecteur contenant les scores de détection des autres concepts. Ce nouveau descripteur est finalement ajouté à d'autres descripteurs déjà existants (de bas niveaux), il est utilisé pour une classification et inclus dans un processus de fusion. La figure 3.2 montre ce principe simplifié. Ce descripteur peut être généré et utilisé une seule fois ou d'une façon itérative.

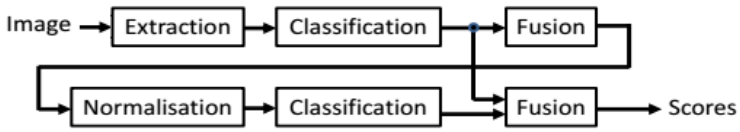

Figure 3.2 Système de détection de concepts avec rétroaction conceptuelle [11]

- Descripteur sémantique utilisant un vecteur de Fisher [12]: dans [12], les auteurs proposent un descripteur qui utilise un vecteur de Fisher comme une alternative aux sacs de mots visuels. On peut avoir plusieurs variantes de ce descripteur, parmi lesquelles, on distingue un descripteur de dimension de 10174 entrainé sur 10174 concepts de ImageNet[6] [4].

3.3. Descripteurs appris

Les descripteurs appris sont générés avec une phase d'apprentissage sur les pixels de l'image. Dans ce type de descripteurs on utilise les informations extraites directement du signal via une phase d'apprentissage (ex. l'utilisation de l'apprentissage profond utilisant les CNN « Convulutional Neural Networks »).

Dans cette approche de génération de descripteurs, on lance l'apprentissage profond, ensuite on utilise les résultats intermédiaires comme un descripteur de haut niveau à intégrer dans un système de détection de concepts.

Les réseaux de neurones convolutionnels profonds (DCNN - Deep Convolutional Neural Networks) ont récemment marqué une amélioration significative dans la classification des images. Cela a été rendu possible par une conjonction de plusieurs facteurs, comme par exemple : la découverte des approches qui donnent des réseaux profonds efficaces, l'utilisation de couches convolutionnelles, la disponibilité d'architectures parallèles très puissantes (GPU), la découverte des approches sur la façon exacte d'organiser ces réseaux pour cette tâche (classification des images), et la disponibilité d'une quantité énorme de données proprement annotées [4].

La disponibilité d'un grand nombre d'exemples d'images pour un très grand nombre de concepts était vraiment cruciale, car les DCNN ont vraiment besoin d'une telle quantité de données d'apprentissage pour être réellement efficaces. L'augmentation de données (ex. des

[6]Projet ImageNet : c'est un effort de recherche en cours pour fournir aux chercheurs du monde entier une base de données d'images facilement accessible.

corpus multiples d'échantillons d'apprentissage) peut également aider, mais aussi uniquement lorsqu'une quantité énorme de données est déjà disponible. Une telle quantité de données d'apprentissage est actuellement disponible uniquement avec ImageNet qui correspond à un seul type d'application et uniquement pour les images fixes. Il existe de nombreuses autres collections annotées mais avec un nombre beaucoup plus petit de concepts et/ou beaucoup moins d'exemples. Le fait de tenter d'entrainer des DCNN sur de telles données conduit généralement à des résultats moins bons que ceux obtenus à l'aide des descripteurs classiques (de bas niveau) combinés à des méthodes d'apprentissage classiques (typiquement SVM) [4].

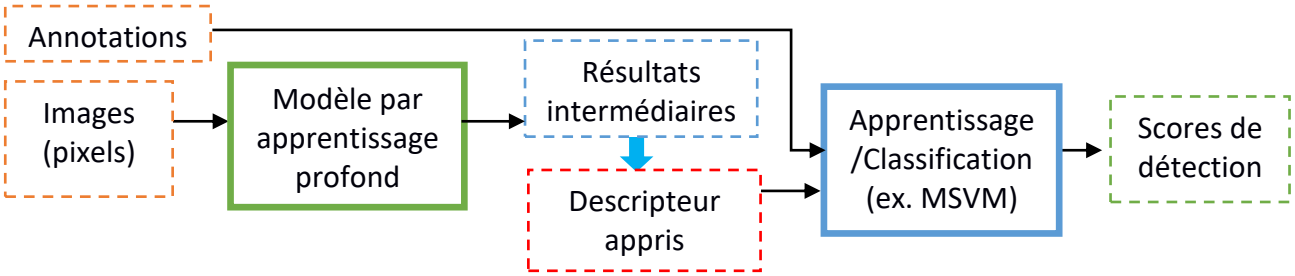

Figure 3.3 Utilisation d'un descripteur appris dans un système de détection de concepts

Dans ce qui suit on va citer quelques travaux utilisant les descripteurs appris :

- Descripteur appris utilisant les DCNN suivant l'architecture de Krizhevsky [13]: c'est une approche qui a utilisé les DCNN pour la classification de 1.2 millions d'images de haute résolution dans un défi d'ImageNet en 2010 [13].

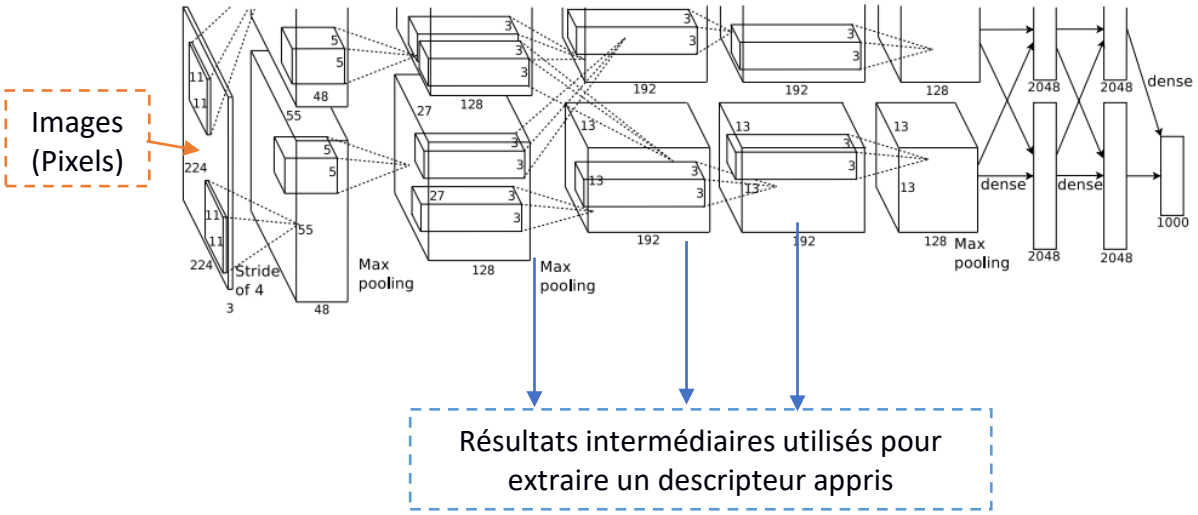

Figure 3.4 Illustration de l'architecture de CNN de Krizhevsky [13]

- Descripteur appris qui correspond aux trois dernières couches cachées [4]: dans ce travail, les auteurs ont utilisé les résultats qui correspondent aux trois dernières couches cachées du DCNN suivant l'architecture de Krizhevsky [4].

3.4. Conclusion

A travers ce chapitre nous avons pu décrire les différents descripteurs d'images de haut niveau qui peuvent être utilisés dans un système de détection de concepts. Dans ce qui suit nous allons présenter nos différentes approches utilisées dans la détection de multi-concepts, utilisant plusieurs descripteurs, avec une présentation et une discussion des résultats.

Chapitre 4

Contributions et expérimentations

4.1. Introduction

Dans ce chapitre nous allons présenter nos contributions, en décrivant les approches utilisées, ensuite nous allons détailler les différentes données utilisées et les expérimentations faites sur ces données, en appliquant plusieurs approches et plusieurs descripteurs. En terminant avec une discussion des résultats d'évaluation obtenus.

4.2. Description des approches

4.2.1. Descripteurs sémantiques utilisant la rétroaction conceptuelle

Nous avons fait l'exploitation d'une version dérivée de l'approche de rétroaction conceptuelle évoquée dans le chapitre 3. Nous rappelons que cette dernière exploite la détection de concepts singuliers utilisant un descripteur initial afin de construire un descripteur sémantique. Ce nouveau descripteur de haut niveau est ensuite utilisé dans une nouvelle étape d'apprentissage pour la détection de multi-concepts, comme illustré dans l'exemple ci-après.

Exemple : Afin de détecter un multi-concept cible (c-1, c-2, …, c-N), nous allons détecter en premier temps un ensemble de concepts singuliers, qui ne contient pas nécessairement les concepts composant le multi-concept cible. On suppose que cet ensemble est composé des concepts singuliers suivants : concept-1, concept-2, …, concept-M. La détection du multi-concept est réalisée de la manière décrite dans la Figure 4.1.

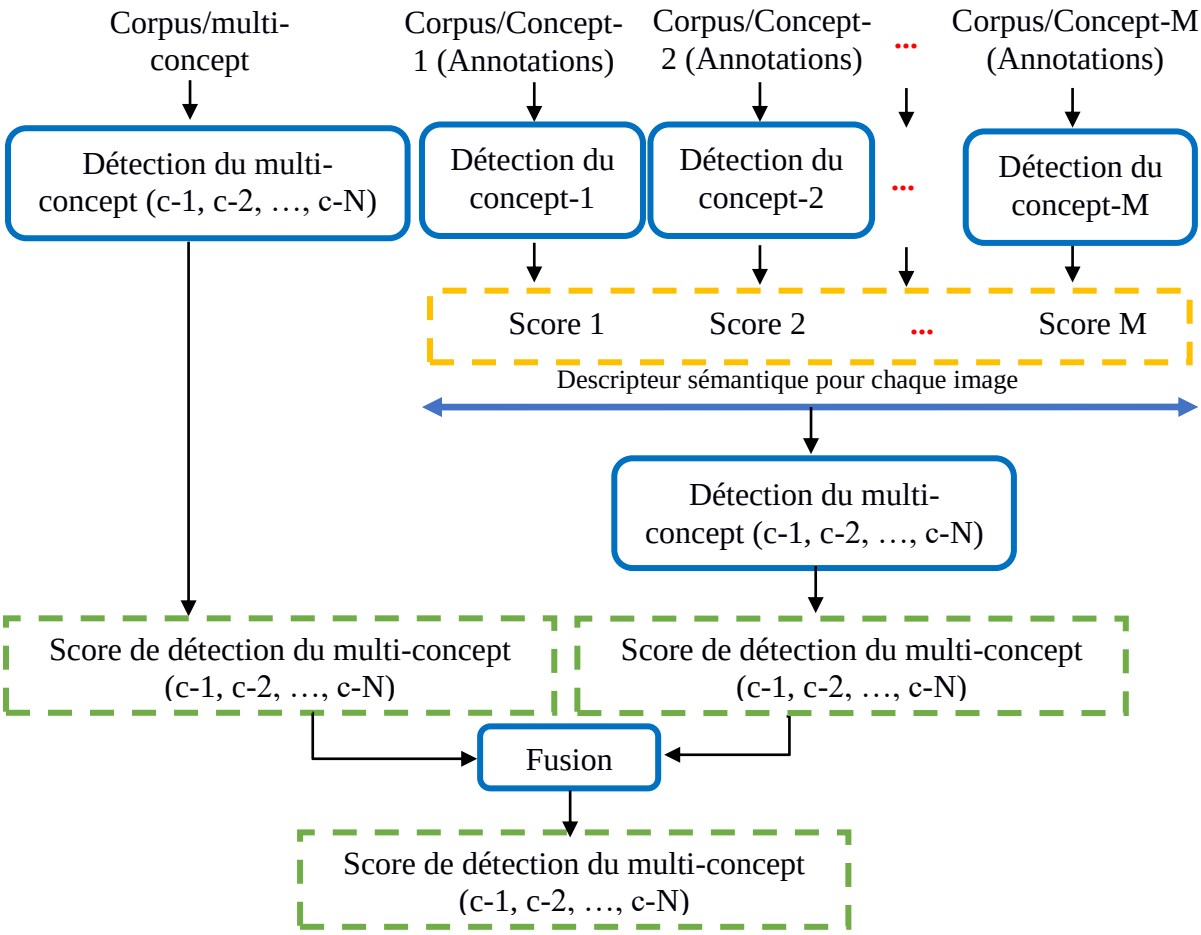

Figure 4.1 Utilisation de la détection des concepts singuliers pour détecter un multi-concept

4.2.2. Descripteurs appris utilisant l'apprentissage profond

Nous avons utilisé des descripteurs appris pour la détection de multi-concept, nous avons opté pour le choix de l'apprentissage profond avec les DCNN pour leurs bons résultats dans les travaux antérieurs de l'état de l'art. D'où nous avons utilisé un modèle d'un DCNN appris sur un corpus de données qui contient une grande masse d'images, afin d'extraire des descripteurs appris pour le corpus de données que nous avons utilisé, en prenant les informations intermédiaires générées sur les images utilisant ce DCNN. Le principe d'utilisation de ce descripteur est montré dans la Figure 4.2.

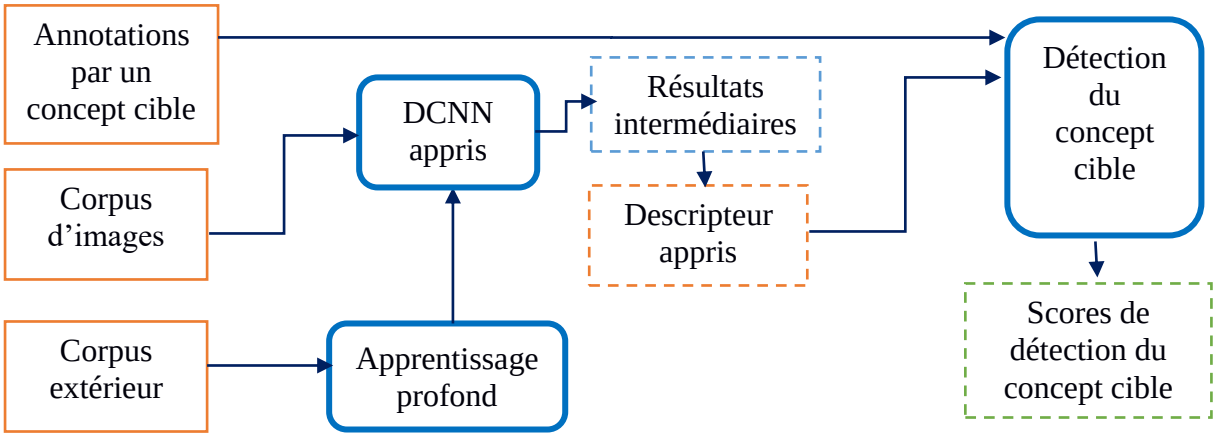

Figure 4.2 Utilisation d'un descripteur appris dans un système de détection de concept

4.3. Données et expérimentations

4.3.1. Corpus Pascal-VOC 2012

Afin de tester et évaluer nos approches utilisées pour la détection de multi-concepts dans les images, nous avons mené notre évaluation sur le corpus Pascal-VOC 2012. Concernant les multi-concepts nous avons considéré les cas de paires et triplet de concepts, parce qu'il n'y a pas une disponibilité de données annotées par des multi-concepts pour des groupes composés d'un plus grand nombre de concepts. Dans ce corpus il y a une liste principale de 20 concepts singuliers. Nous avons retenu 60 paires et 45 triplets de concepts formés à partir de ces concepts singuliers. Les annotations des images par les concepts singuliers ont été fournies dans le même corpus, par contre nous avons généré les annotations par les multi-concepts utilisant la méthode de génération d'un modèle de concepts multiples par l'intersection des annotations des concepts comme ce qui est décrit dans le premier chapitre, Ce qui a donné, comme prévu, un petit nombre d'exemples positifs. Le nombre d'exemples positifs pour les paires de concepts varie entre 10 et 380, alors que ce nombre varie entre 7 et 98 pour les triplets de concepts. Nous avons évalué nos différentes approches proposées en termes de précision moyenne (MAP). Le *Tableau 1* montre les détails du corpus et des concepts.

Tableau 1 Corpus Pascal-VOC 2012

Nombres des images		**Nombres des concepts**		
Apprentissage	Validation	Singuliers	Paires	Triplets
5717	5823	20	60	45

4.3.2. Descripteurs de bas niveau

Nous avons utilisé pour la description du contenu des images deux type de descripteurs de bas niveau : histogramme de couleur, descripteur en sac de mots visuels utilisant SIFT.

- rgbHist8bins : dans ce descripteur nous avons calculer l'histogramme de couleurs, dans l'espace de couleurs RGB, en découpant les valeurs possibles d'intensité des pixels à 8 bins « intervalles ».
- sift1024 : dans ce descripteur nous avons créé une description en sac de mots visuels des SIFT de dimension de 1024, utilisant la méthode de clustering K-means.

4.3.3. Multi-SVM pour la détection de concepts

Nous avons utilisé les Multi-SVM comme méthode d'apprentissage et de classification, pour la détection des concepts, parce qu'ils donnent de bons résultats face au problème de classes déséquilibrées, comme ce qui montré dans le premier chapitre. Nous avons appliqué ces Multi-SVM pour la détection des différents types de concept du corpus Pascal-VOC 2012, utilisant les différents types de descripteurs de bas niveau, sémantiques et appris.

4.3.4. Fusion tardive

Nous avons utilisé l'approche de la fusion tardive qui est décrite dans le premier chapitre. Nous avons fait la fusion des résultats de détection pour les différents types de concepts, utilisant les deux descripteurs de bas niveau (rgbHist8bins et sift1024), nous avons appelé le résultat « fusionHS ».

4.3.5. Détection de multi-concepts

Nous avons utilisé pour la détection de multi-concepts les deux approches décrites dans le premier chapitre. Nous avons utilisé première approche « learnMulti » qui consiste à faire la détection sur un modèle de concepts multiples, en générant des modèles de paires et de triplets de concepts utilisant les annotations des concepts singulier, et faisant la détection utilisant les Multi-SVM sur ces modèles. Nous avons utilisé la deuxième approche « singleFus » qui consiste à utiliser les résultats de détection des concepts singuliers et les fusionner afin d'obtenir les résultats pour les multi-concept, en utilisant la détection des concepts singuliers par Multi-SVM, et faisant la fusion des résultats de détection pour avoir des résultats de détection des paires et des triplets de concepts. Nous avons appliqué ces deux approches utilisant les différents types de descripteurs de bas niveau, sémantiques et appris.

4.3.6. Descripteurs sémantiques

Nous avons utilisé pour la détection de multi-concepts, en outre des descripteurs de bas niveau (rgbHist8bins, sift1024) et la fusion tardive fusionHS, un autre descripteur sémantique qui est de haut niveau. Nous avons fait plusieurs expérimentations pour la détection de concepts singuliers utilisant les deux descripteurs de bas niveau (rgbHist8bins, sift1024) et la fusionHS, ensuite nous avons pris les meilleurs résultats dans cette tâche, afin de construire un descripteur sémantique par l'approche de rétroaction conceptuelle utilisant ces résultats. Nous avons fait la détection des paires et des triplets de concepts avec ce descripteur sémantique.

4.3.7. Descripteurs appris

Nous avons utilisé le modèle de DCNN GoogleNet [14], généré par apprentissage profond utilisant la collection d'images d'ILSVRC2012[7], pour extraire des informations intermédiaires de la couche « pool5/7x7_s1 » de ce modèle sur les images de notre corpus (Pascal-VOC 2012), ensuite nous avons construit un descripteur appris de dimension de 1024 valeurs pour chaque image. Nous avons utilisé ce descripteur pour la détection des différents types de concepts.

Nous avons appliqué aussi l'approche d'extraction de descripteur sémantique utilisant la rétroaction conceptuelle, avec ce descripteur appris (GoogleNet) comme descripteur initial.

4.4. Résultats et discussion

Après avoir faire nos expérimentations, et évaluer les performances des différentes approches utilisées. Nous avons récapitulé les résultats d'évaluation de chaque approche en terme de MAP dans le *Tableau 2*.

Tableau 2 Résultats d'évaluation des performances des approches en terme de MAP

			learnMulti		**SingleFus**	
	Descripteur/Concept	*Singulier*	*Paire*	*Triplet*	*Paire*	*Triplet*
Descripteurs de bas niveau	*rgbHist8bins*	0.125	0.0145	0.0033	0.0203	0.0088
	sift1024	0.1836	0.0197	0.0056	0.0273	0.0103
	fusionHS	0.2283	0.0276	0.0061	0.0371	0.0152
Descripteurs de haut niveau	*Sem/fusionHS*	0.2666	0.0453	0.0169	0.0467	0.0199
	GoogleNet	0.8081	0.1802	0.0567	0.2707	0.0797
	Sem/GoogleNet	0.8004	0.1898	0.0583	0.2375	0.0652

[7] http://www.image-net.org/challenges/LSVRC/2012/

4.4.1. Détection de concepts singuliers

Nous avons testé toutes nos approches en premier temps sur les concepts singuliers, pour bien voir l'efficacité de ces approches, et si elles donnent des résultats qui correspondent bien à ce qui est montré dans l'état de l'art. La Figure 4.3 montre les valeurs d'évaluation en terme de MAP pour la détection des concepts singuliers utilisant les différents descripteurs.

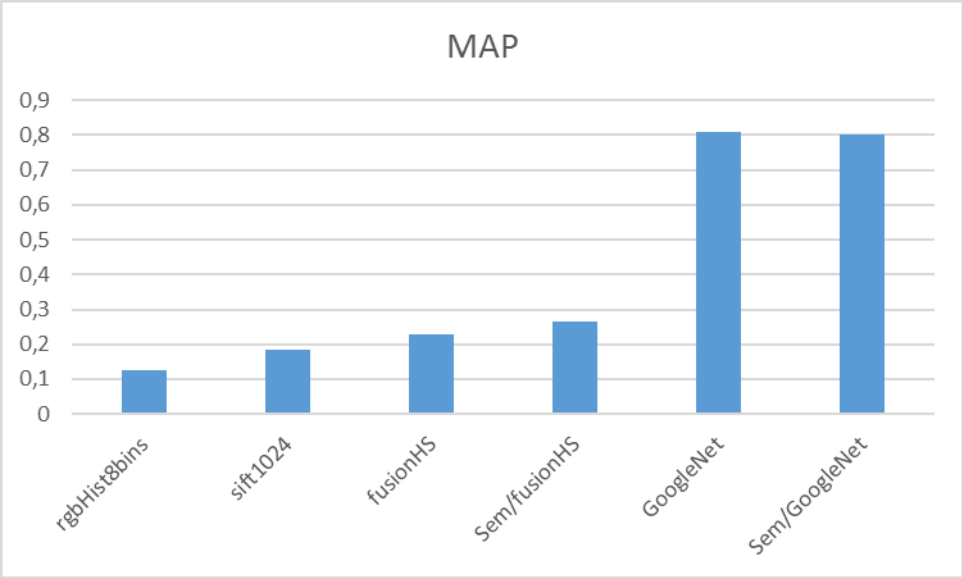

Figure 4.3 Performances des différents descripteurs pour la détection de concepts singuliers en terme de MAP

- **Descripteurs de bas niveau**

Nous remarquons pour les descripteurs de bas niveau, que l'utilisation des SIFT donne des résultats meilleurs que ceux de l'utilisation des histogrammes de couleurs, et que la fusion de ces deux descripteurs donne encore de meilleurs résultats que ceux de l'utilisation des SIFT. Nous avons pour l'utilisation de fusionHS par rapport à sift1024 un pourcentage de gain relatif[8] de +24%.

- **Descripteurs de haut niveau**

Nous remarquons pour les descripteurs de haut niveau, que leur utilisation donne des performances significativement bien meilleures que celles de l'utilisation des descripteurs de bas niveau dans la détection de concepts singuliers, ce qui est montré dans l'état de l'art. Nous avons pour l'utilisation de Sem/fusionHS par rapport à fusionHS un pourcentage de gain relatif de +16,77%. Nous remarquons aussi que pour l'utilisation de GoogleNet par rapport à fusionHS on obtient un pourcentage de gain relatif de +254%, ce qui montre que le descripteur appris donne de très bons résultats. L'utilisation de Sem/GoogleNet par rapport à GoogleNet donne un pourcentage de gain relatif de -0.01%, ce qui revient à la situation où les résultats d'utilisation de la rétroaction conceptuelle atteignent une sorte de convergence après un certain nombre d'itérations [15].

La Figure 4.4 montre les résultats d'évaluation en terme de AP pour l'utilisation de fusionHS, Sem/fusionHS et GoogleNet, dans la détection de certains concepts singuliers.

[8] Gain relatif = ((Nouvelle MAP – Ancienne MAP) / (Ancienne MAP)) x 100

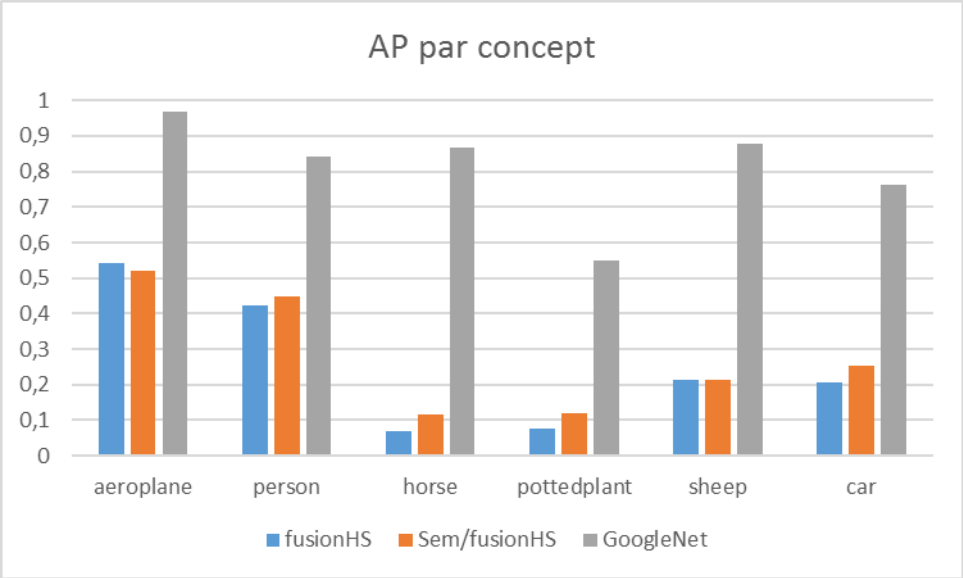

Figure 4.4 Résultats de détection de quelques concepts avec trois différents types de descripteur en terme de AP

Nous remarquons que le descripteur sémantique Sem/fusionHS donne généralement de meilleurs résultats que fusionHS, malgré qu'il est moins performant pour certains concepts comme pour le concept « aeroplane ». Nous remarquons aussi que le descripteur appris GoogleNet donne toujours les meilleures performances par rapport aux autres descripteurs pour tous les concepts.

4.4.2. Détection de multi-concepts

Comme en premier temps sur les concepts singuliers, nous avons testé toutes nos approches deuxièmement dans la détection de multi-concepts. Nous avons testé les performances des deux approches de détection de multi-concept (learnMulti et singleFus), pour bien voir l'efficacité de toutes les approches, et si elles donnent des améliorations de performances comme dans la détection de concepts singuliers. La Figure 4.5 montre les valeurs d'évaluation en terme de MAP pour la détection des paires et des triplets de concepts utilisant les différents descripteurs.

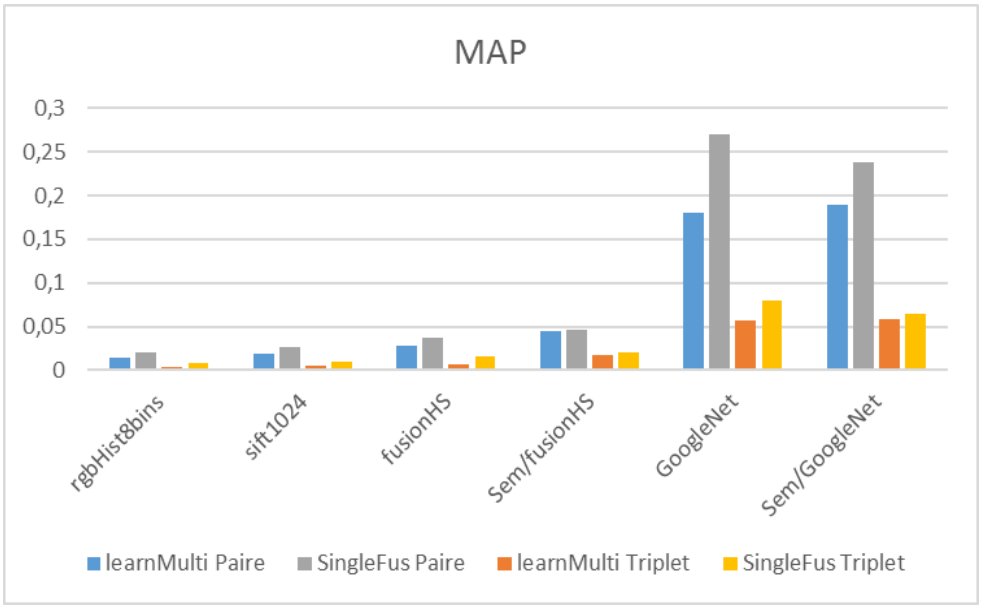

Figure 4.5 Performances des différents descripteurs et approches pour la détection de multi-concepts en terme de MAP

4.4.2.1. Détection par modèles de multi-concepts « learnMulti »

Dans la détection de multi-concepts, nous avons testé les performances des différents descripteurs utilisant l'approche de détection par modèles de concepts multiples.

- **Descripteurs de bas niveau**

Nous remarquons pour les descripteurs de bas niveau, que l'utilisation des SIFT donne de meilleures performances que celles de l'utilisation des histogrammes de couleurs, et que la fusion de ces deux descripteurs donne encore de meilleures performances que celles de l'utilisation des SIFT. Nous avons pour l'utilisation de fusionHS par rapport à sift1024 dans la détection de paires de concepts un pourcentage de gain relatif de +40,1% et dans la détection de triplets de concepts un pourcentage de gain relatif de +9%.

- **Descripteurs de haut niveau**

Nous remarquons pour les descripteurs de haut niveau, que leur utilisation donne des performances significativement bien meilleures que celles de l'utilisation des descripteurs de bas niveau dans la détection de multi-concepts, ce qui montre qu'ils donnent les meilleurs résultats dans la détection de multi-concepts tout comme dans la détection de concepts singuliers. Nous avons pour l'utilisation de Sem/fusionHS par rapport à fusionHS dans la détection de paires de concepts un pourcentage de gain relatif de +64,41% et dans la détection de triplets de concepts un pourcentage de gain relatif de +177%. Nous remarquons aussi que pour l'utilisation de GoogleNet par rapport à fusionHS dans la détection de paires de concepts on obtient un pourcentage de gain relatif de +553% et pout la détection de triplets de concepts un pourcentage de gain relatif de +829%, ce qui montre que le descripteur appris donne de très bons résultats dans la détection de multi-concepts meilleurs que ceux des descripteurs de bas niveau ainsi que ceux de leur fusion. L'utilisation de Sem/GoogleNet par rapport à

GoogleNet dans la détection de paires de concepts donne un pourcentage de gain relatif de +5,32% et dans la détection de triplets de concepts un pourcentage de gain relatif de +2,82%, ce qui montre que l'utilisation d'un descripteur sémantique utilisant la rétroaction conceptuelle extrait à partir d'un descripteur appris peut améliorer significativement les performances de détection de multi-concepts.

La Figure 4.6 montre les résultats d'évaluation en terme de AP pour l'utilisation de fusionHS, Sem/fusionHS, GoogleNet et Sem/GoogleNet dans la détection de certains paires et triplets de concepts, utilisant l'approche learnMulti.

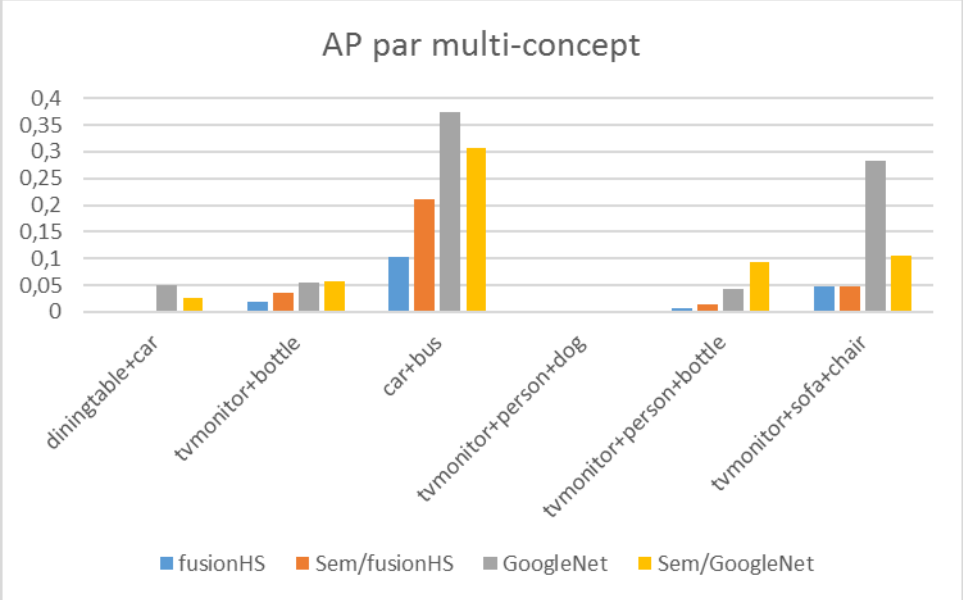

Figure 4.6 Résultats de détection de quelques multi-concepts avec quatre différents descripteurs utilisant learnMulti en terme de AP

Nous remarquons que le descripteur sémantique Sem/fusionHS donne généralement de meilleurs résultats que fusionHS. Nous remarquons aussi que le descripteur appris GoogleNet donne de très bons résultats. En effet, pour certains multi-concepts, la performance est meilleure que Sem/GoogleNet qui a la meilleure performance en terme de MAP.

Nous remarquons que pour certains concepts qui ont très peu d'exemples positifs dans le corpus d'apprentissage, la performance de la détection a été faible en utilisant les descripteurs qui donnent généralement les meilleurs résultats pour la détection de multi-concepts, ce qui est le cas pour le tri-concept « tvmonitor+person+dog ».

4.4.2.2. Détection par fusion « singleFus »

Dans la détection de multi-concepts, nous avons testé en outre des performances des différents descripteurs utilisant l'approche de détection par modèles de concepts multiples, leurs performances utilisant l'approche de détection de multi-concepts par fusion de détecteurs de concepts singuliers. Les performances de cette deuxième approche sont liées aux performances de détection de concept singuliers.

- **Descripteurs de bas niveau**

Nous remarquons pour les descripteurs de bas niveau, que l'utilisation des SIFT donne de meilleures performances que ceux de l'utilisation des histogrammes de couleurs, et que la fusion de ces deux descripteurs donne encore de meilleures performances que celles de l'utilisation des SIFT. Nous avons pour l'utilisation de fusionHS par rapport à sift1024 dans la détection de paires de concepts un pourcentage de gain relatif de +35,89% et dans la détection de triplets de concepts un pourcentage de gain relatif de +47,57%.

- **Descripteurs de haut niveau**

Nous remarquons pour les descripteurs de haut niveau, que leur utilisation donne des performances significativement bien meilleures que celles de l'utilisation des descripteurs de bas niveau dans la détection de multi-concepts utilisant les deux approches (learnMulti et singleFus). Utilisant l'approche de fusion des détecteurs de de concepts singuliers, nous avons pour l'utilisation de Sem/fusionHS par rapport à fusionHS dans la détection de paires de concepts un pourcentage de gain relatif de +27,87% et dans la détection de triplets de concepts un pourcentage de gain relatif de +30,92%. Nous remarquons aussi que pour l'utilisation de GoogleNet par rapport à fusionHS dans la détection de paires de concepts on obtient un pourcentage de gain relatif de +629,64% et pout la détection de triplets de concepts un pourcentage de gain relatif de +424,34%, ce qui montre que le descripteur appris donne de très bons résultats dans la détection de multi-concepts utilisant les deux approches (learnMulti ou singleFus) meilleurs que ceux des descripteurs de bas niveau ainsi que ceux de leur fusion. L'utilisation de Sem/GoogleNet par rapport à GoogleNet dans la détection de paires de concepts utilisant l'approche singleFus donne un pourcentage de gain relatif de -12,26% et dans la détection de triplets de concepts un pourcentage de gain relatif de -18,19%, ce qui montre que l'utilisation d'un descripteur sémantique utilisant la rétroaction conceptuelle extrait à partir d'un descripteur appris, ne donne pas toujours de meilleures performances que celles de l'utilisation du descripteur appris directement. Nous remarquons aussi que lorsque les performances de détection de concepts singuliers diminuent, alors les performances de détection de multi-concepts utilisant singleFus diminuent aussi, parce que cette approche utilise les résultats de détection de concept singuliers.

Nous remarquons que pour la majorité des descripteurs utilisés, que l'utilisation de l'approche de fusion de détecteurs de concepts singuliers (singleFus) donne de meilleures performances que celles de l'utilisation de l'approche de détection par modèles de concepts multiples.

La Figure 4.7 montre les résultats d'évaluation en terme de AP pour l'utilisation de fusionHS, Sem/fusionHS, GoogleNet et Sem/GoogleNet dans la détection de certains paires et triplets de concepts, utilisant l'approche singleFus.

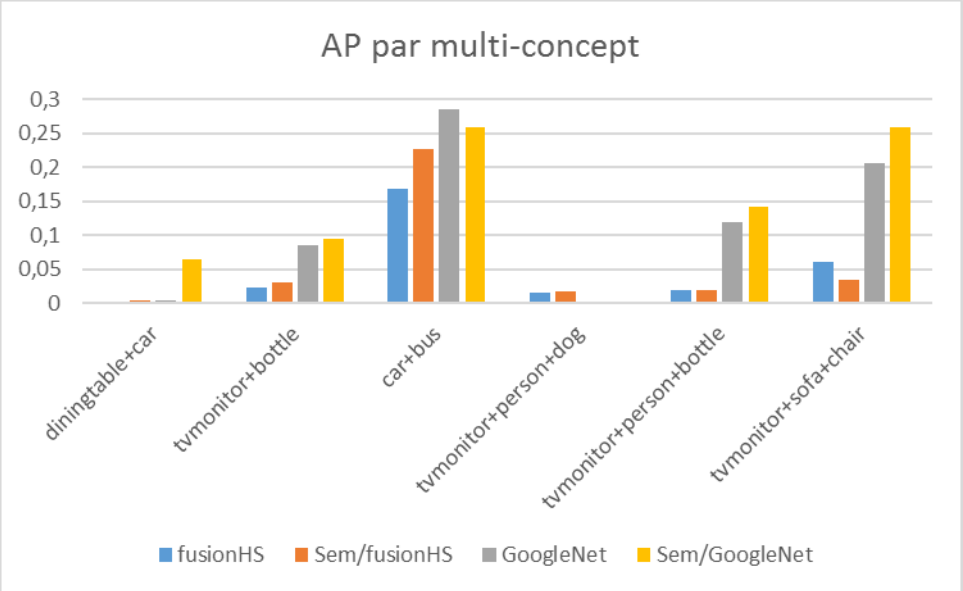

Figure 4.7 Résultats de détection de quelques multi-concepts avec quatre différents descripteurs utilisant singleFus en terme de AP

4.5. Conclusion

A travers ce chapitre nous avons pu décrire nos différentes approches utilisées dans un système de détection de concepts, pour la détection de multi-concepts. Nous avons détaillé ensuite les différentes expérimentations faites afin d'évaluer nos approches, utilisant les corpus Pascal-VOC 2012. Nous avons terminé par une présentation des résultats et une discussion. Dans ce qui suit nous allons présenter les différentes étapes de conception et d'implémentation de notre système de détection de concept.

Chapitre 5

Conception et implémentation

5.1. Introduction

Dans ce chapitre nous abordons les aspects de conception de notre système, en montrant les différents diagrammes de conception, qui décrit l'architecture et le fonctionnement de notre système de détection de concepts. Nous entamons ensuite la présentation de l'environnement de travail, en détaillants les différents outils utilisés. Nous allons présenter aussi, les différentes interfaces principales du système.

5.2. Conception

Nous avons modélisé les besoins de notre système de détection de concepts, en se basant sur le langage de modélisation UML « Unified Modeling Language ». Nous avons choisi pour cette modélisation deux diagrammes principaux, le diagramme de cas d'utilisation et le diagramme de classes, qui représentent les besoins fonctionnels ainsi que l'architecture du système.

Nous avons utilisé comme outils de modélisation UML, l'outil StarUML qui est un logiciel de modélisation, qui aide à tracer les différents diagrammes UML. Cet outil donne la possibilité d'exporter les diagrammes réalisés en images de format « png » ou « jpg » afin de les insérés au sein du document. La Figure 5.1 montre l'interface principale de l'outil de modélisation StarUML version 2.8.0.

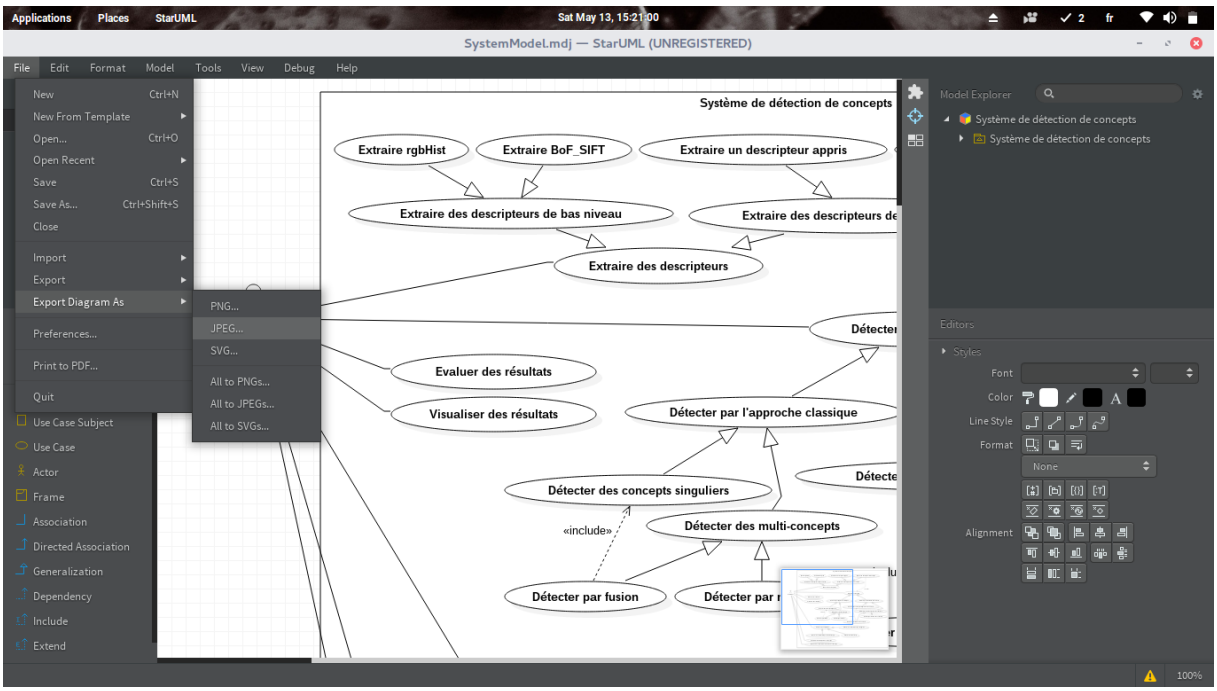

Figure 5.1 Interface principale de StarUML version 2.8.0

5.2.1. Diagramme de cas d'utilisation

Nous avons commencé la modélisation par le diagramme de cas d'utilisation, qui permet de recueillir, d'analyser et d'organiser les besoins, et de recenser les grandes fonctionnalités du système. La Figure 5.2 montre le diagramme de cas d'utilisation de notre système de détection de concepts.

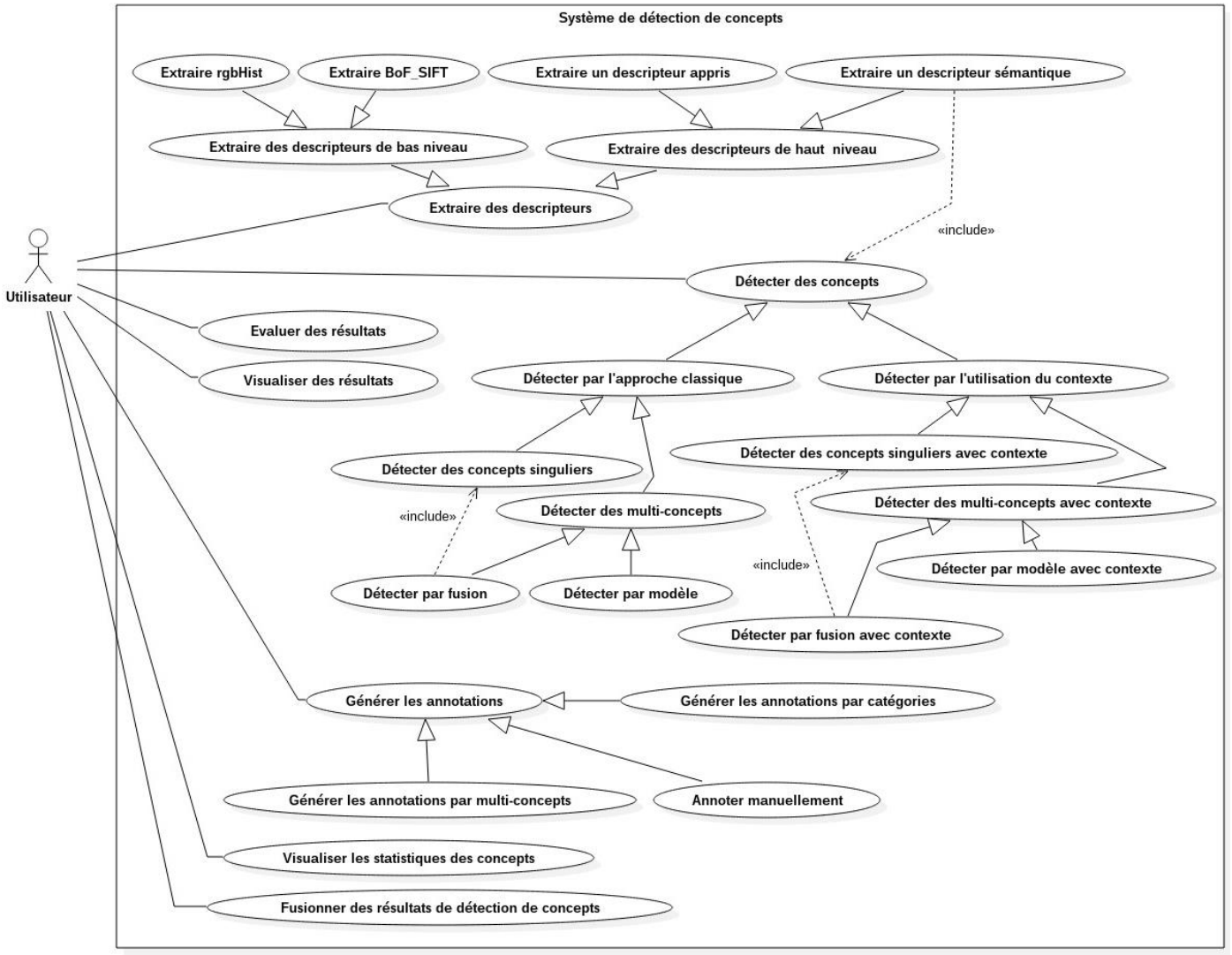

Figure 5.2 Diagramme de cas d'utilisation

5.2.2. Diagramme de classes

Nous avons utilisé le diagramme de classes pour son importance dans la modélisation orientée objet. Alors que le diagramme de cas d'utilisation montre un système du point de vue de l'interaction entre les utilisateurs et le système, le diagramme de classes en montre la structure interne du système. Il permet de fournir une représentation abstraite des objets du système qui vont interagir pour réaliser les cas d'utilisation. Il s'agit d'une vue statique, car on ne tient pas compte du facteur de temps dans le comportement du système. La Figure 5.3 montre le diagramme de classes de notre système de détection de concepts.

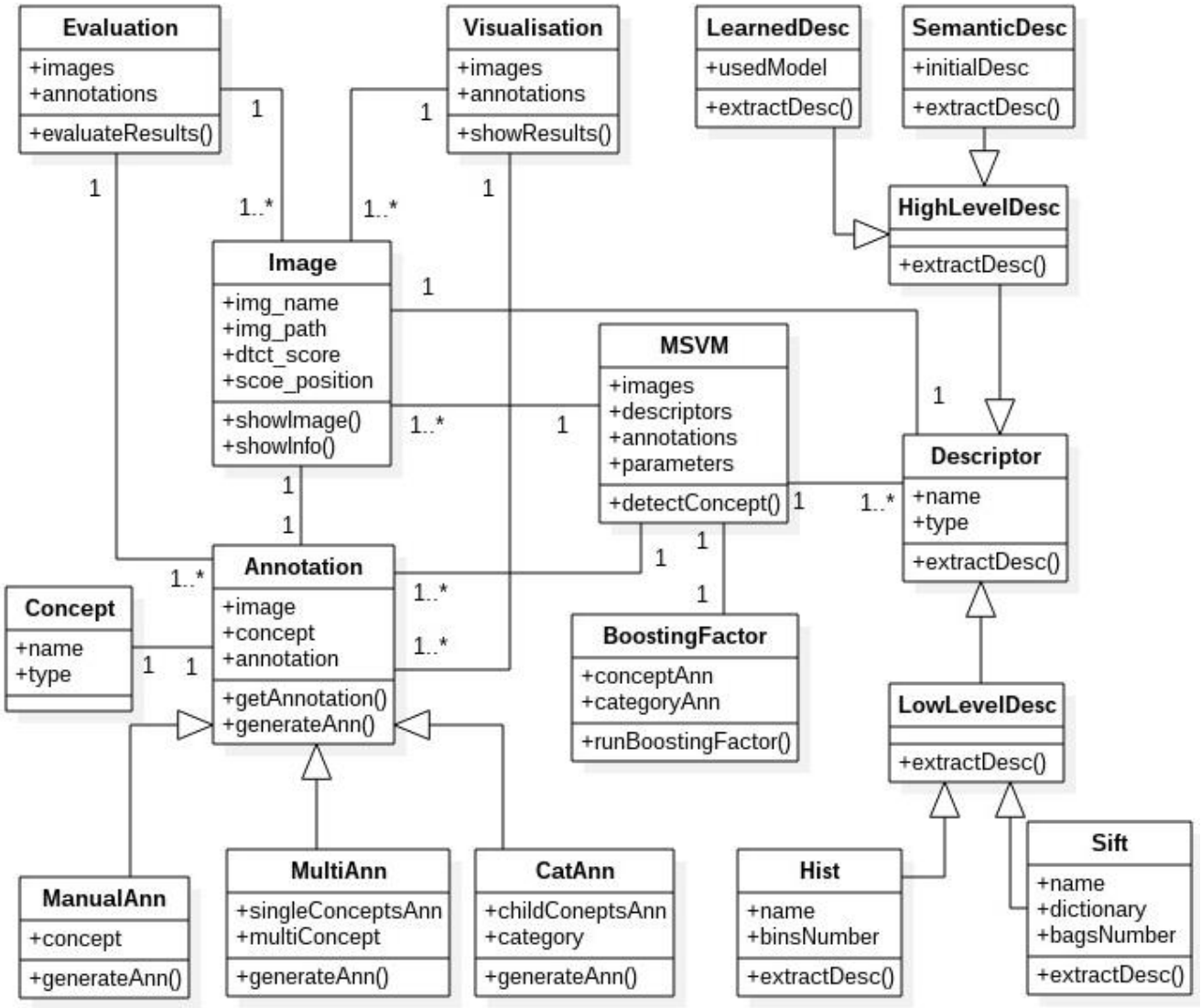

Figure 5.3 Diagramme de classes

5.3. Implémentation

5.3.1. Environnement matériel et logiciel

La réalisation de notre système de détection de concepts, et les différentes expérimentations de nos approches, ont été faites dans l'environnement matériel et logiciel suivant :

- CPU : Intel® Core™ i5-2410M @ 2.30GHz
- RAM : 8 GO
- Système d'exploitation : Linux, distribution Ubuntu 16.04.2 LTS 64-bit
- Outil de modélisation : StarUML version 2.8.0
- Outils de développement : NetBeans IDE 8.2, Sublime Text

5.3.2. Langages et outils de développement

Nous avons utilisé plusieurs langages et outils pour le développement de notre système de détection de concepts. Nous avons utilisé ces langages et outils comme suit :

- Nous avons utilisé le langage de programmation orientée objet Java, utilisant l'outil de développement NetBeans IDE 8.2, pour réaliser les interfaces graphiques de notre système, ainsi que plusieurs traitements.
- Nous avons utilisé les langages de programmation C et C++, utilisant l'outil Sublime Text et la bibliothèque de traitement d'images OpenCV[9], pour extraire les différents descripteurs à partir des images.
- Nous avons utilisé le langage de programmation Python, utilisant l'outil Sublime Text et le framework de l'apprentissage profond Caffe[10] qui est une initiative de l'université de Berkeley, pour extraire les descripteurs profond utilisant un modèle de DCNN appris par une implémentation de Caffe.
- Nous avons utilisé le langage de script Shell « bash », utilisant l'outil Sublime Text, pour lancer plusieurs expérimentations d'optimisation des paramètres de Multi-SVM, d'apprentissage, d'évaluation des résultats et plusieurs autres traitements.

5.3.3. Les interfaces graphiques principales

Dans cette partie nous allons présenter les différentes parties de notre système de détection de concepts, du côté interface graphique, en montrant les interface graphiques principales qui représentent quelques fonctionnalités du système.

- La Figure 5.4 montre l'interface graphique de démarrage du système.

[9] http://opencv.org/
[10] http://caffe.berkeleyvision.org/

Figure 5.4 Interface graphique de démarrage du système

- La Figure 5.5 montre l'interface graphique d'extraction des descripteurs de bas niveau.

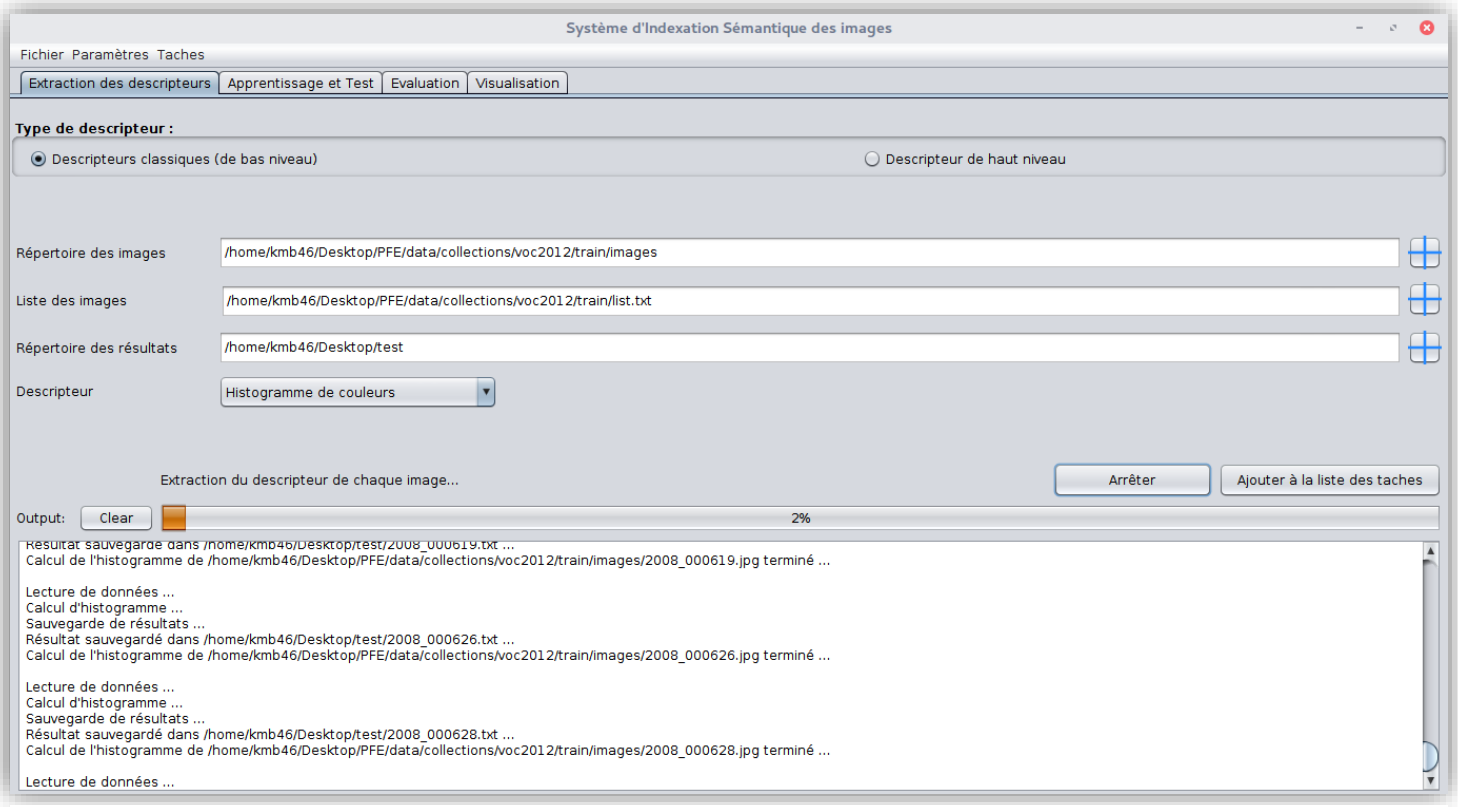

Figure 5.5 Interface graphique d'extraction des descripteurs de bas niveau

- La Figure 5.6 montre l'interface graphique des paramètres des SIFTs.

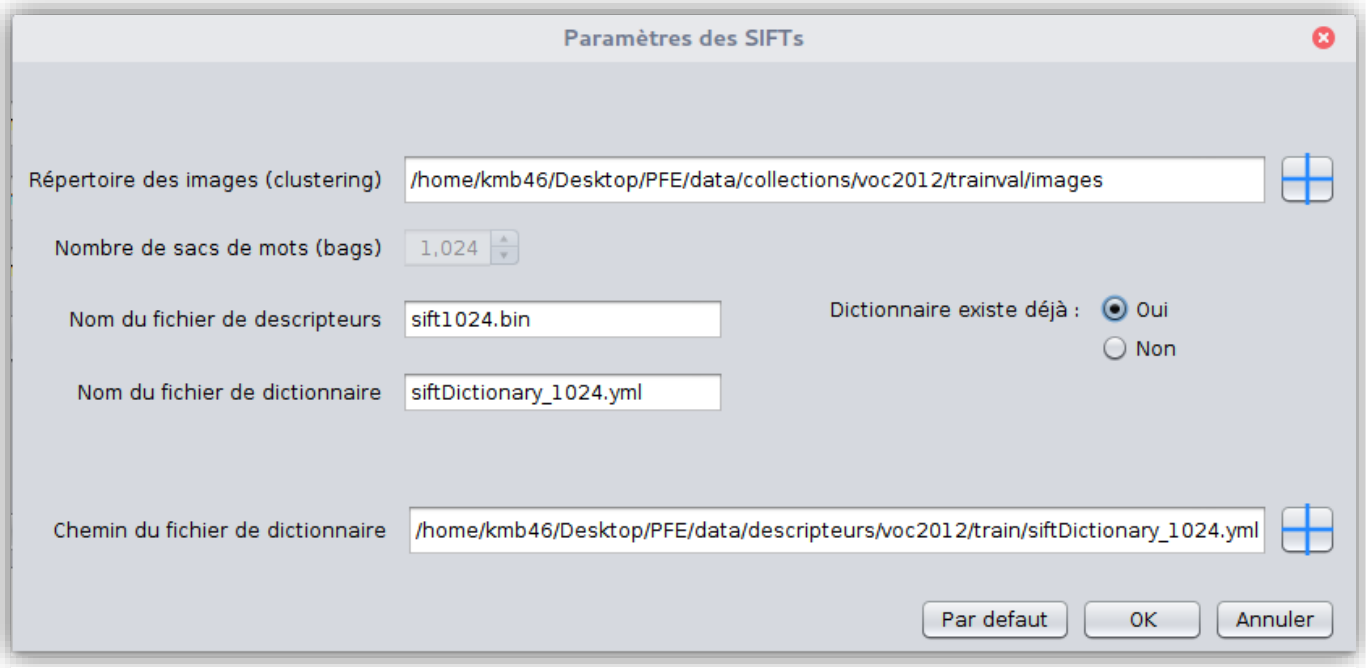

Figure 5.6 Interface graphique des paramètres des SIFTs

- La Figure 5.7 montre l'interface graphique d'extraction des descripteurs sémantiques.

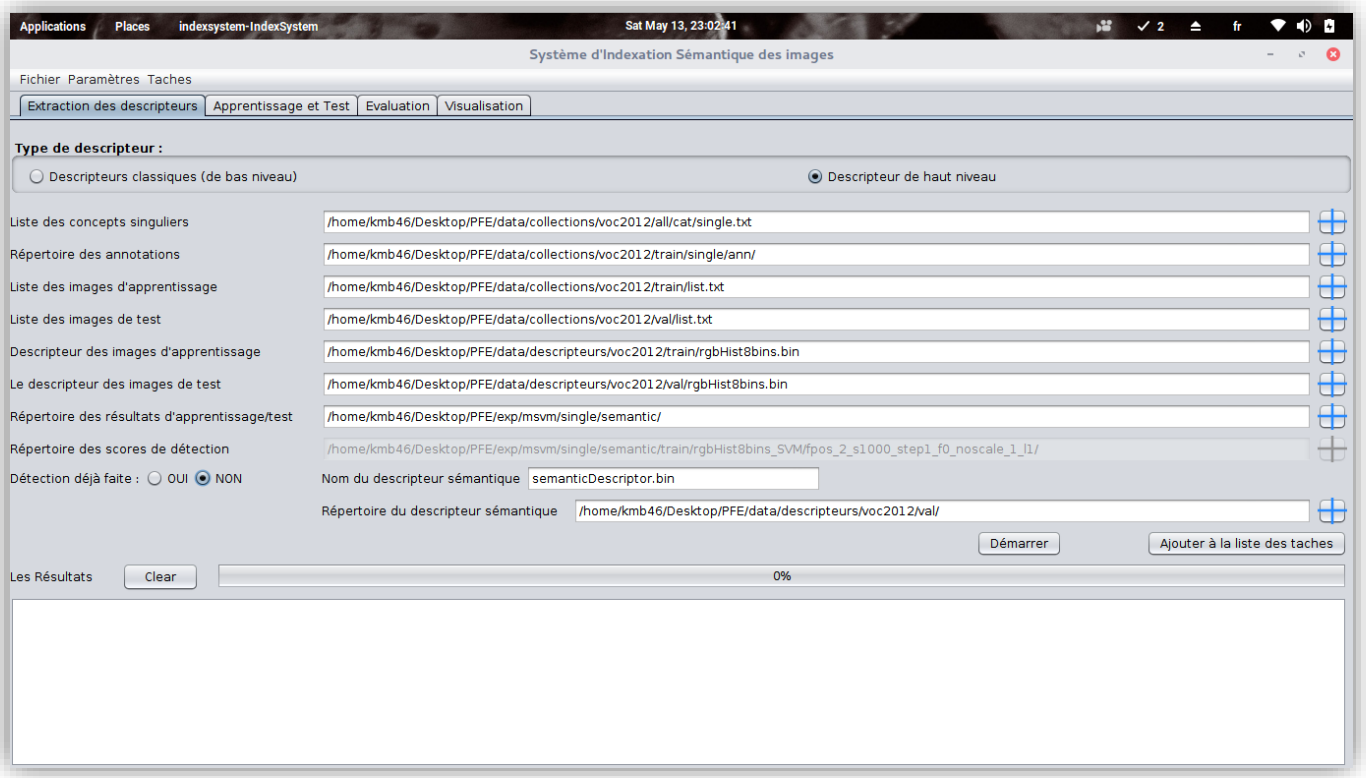

Figure 5.7 Interface graphique d'extraction des descripteurs sémantiques

- La Figure 5.8 montre l'interface graphique des paramètres des Multi-SVM.

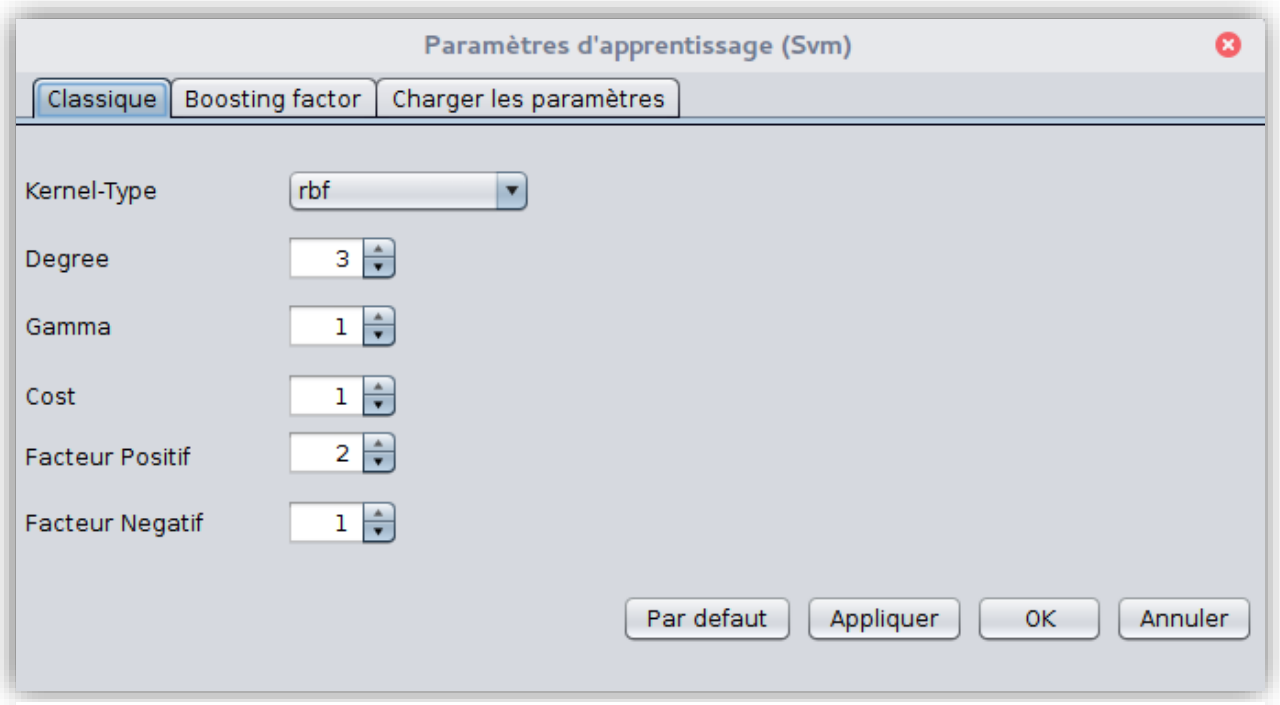

Figure 5.8 Interface graphique des paramètres des Multi-SVM

- La Figure 5.9 montre l'interface graphique de détection des concepts singuliers.

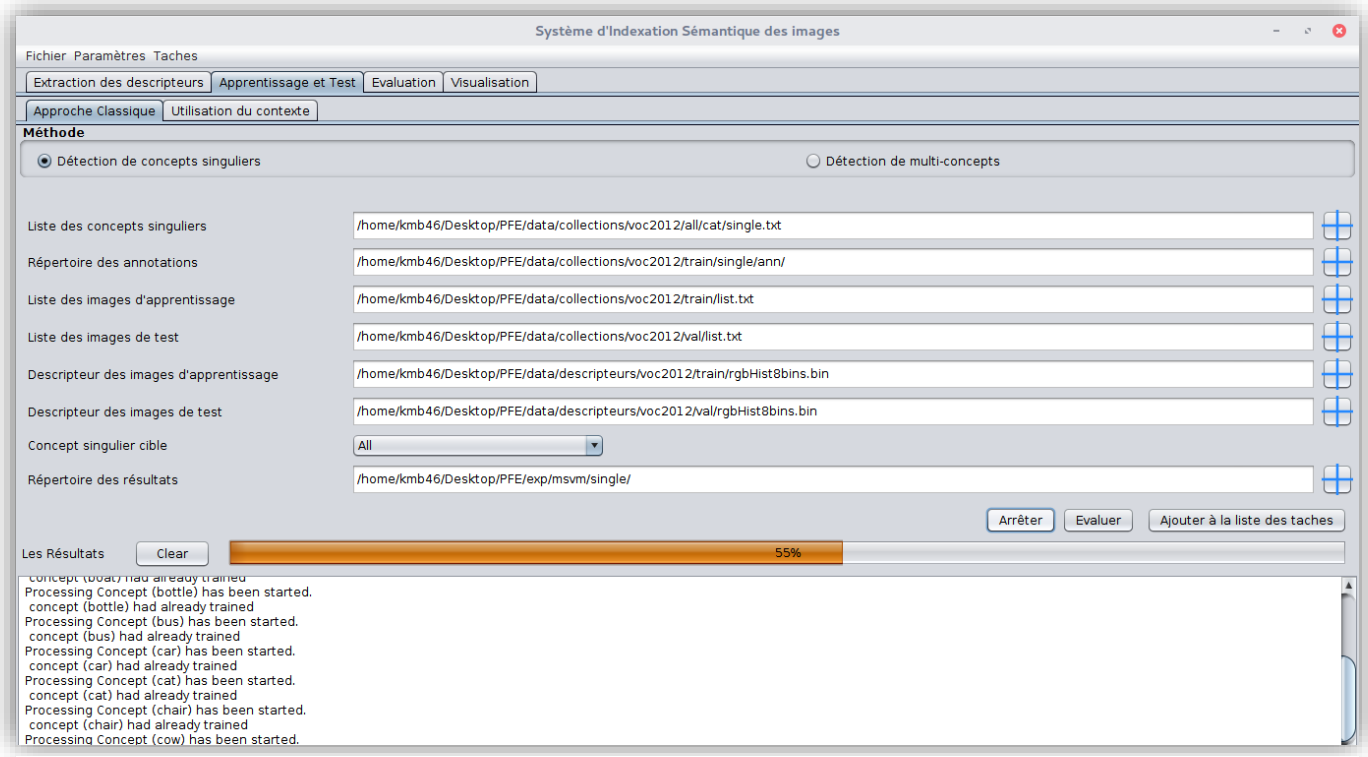

Figure 5.9 Interface graphique de détection des concepts singuliers

- La Figure 5.10 montre l'interface graphique de détection de multi-concepts utilisant l'approche learnMulti.

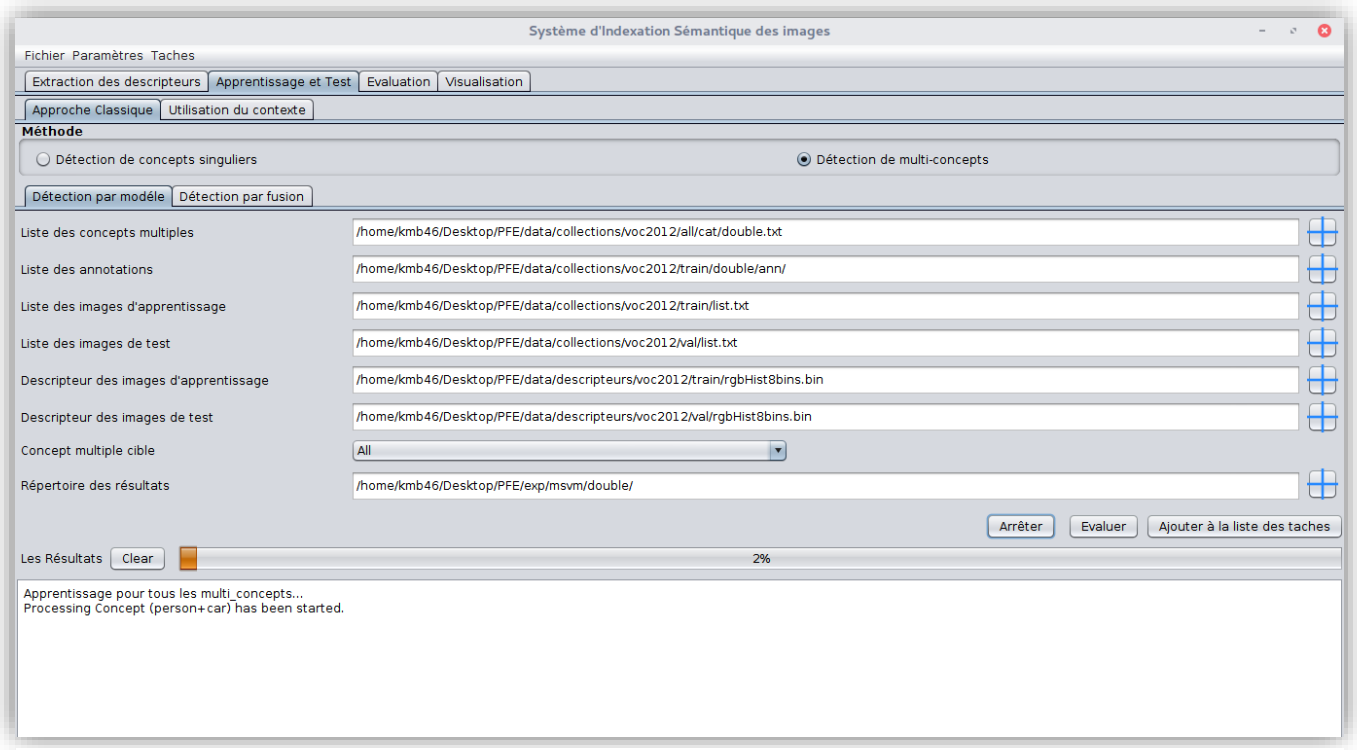

Figure 5.10 Interface graphique de détection de multi-concepts utilisant l'approche learnMulti

- La Figure 5.11 montre l'interface graphique de détection de multi-concepts utilisant l'approche singleFus.

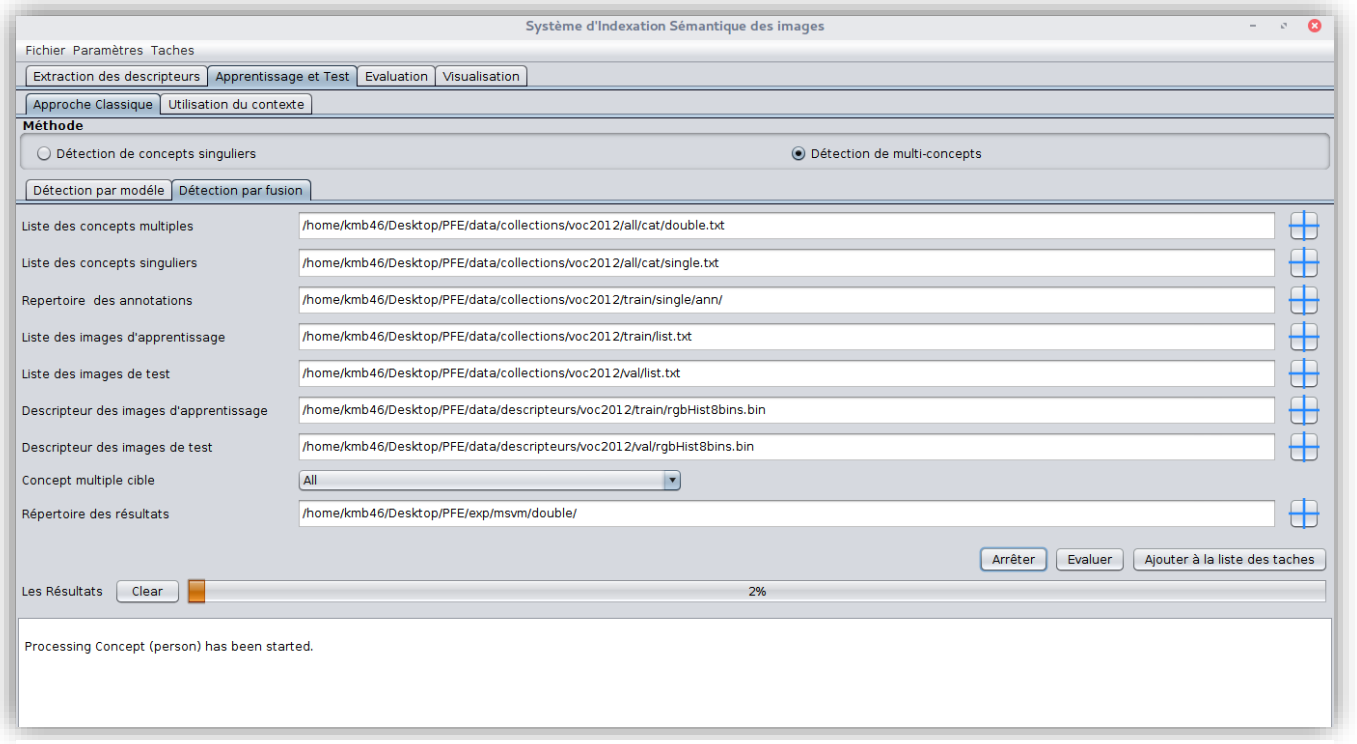

Figure 5.11 Interface graphique de détection de multi-concepts utilisant l'approche singleFus

- La Figure 5.12 montre l'interface graphique d'évaluation des résultats de détection de concepts.

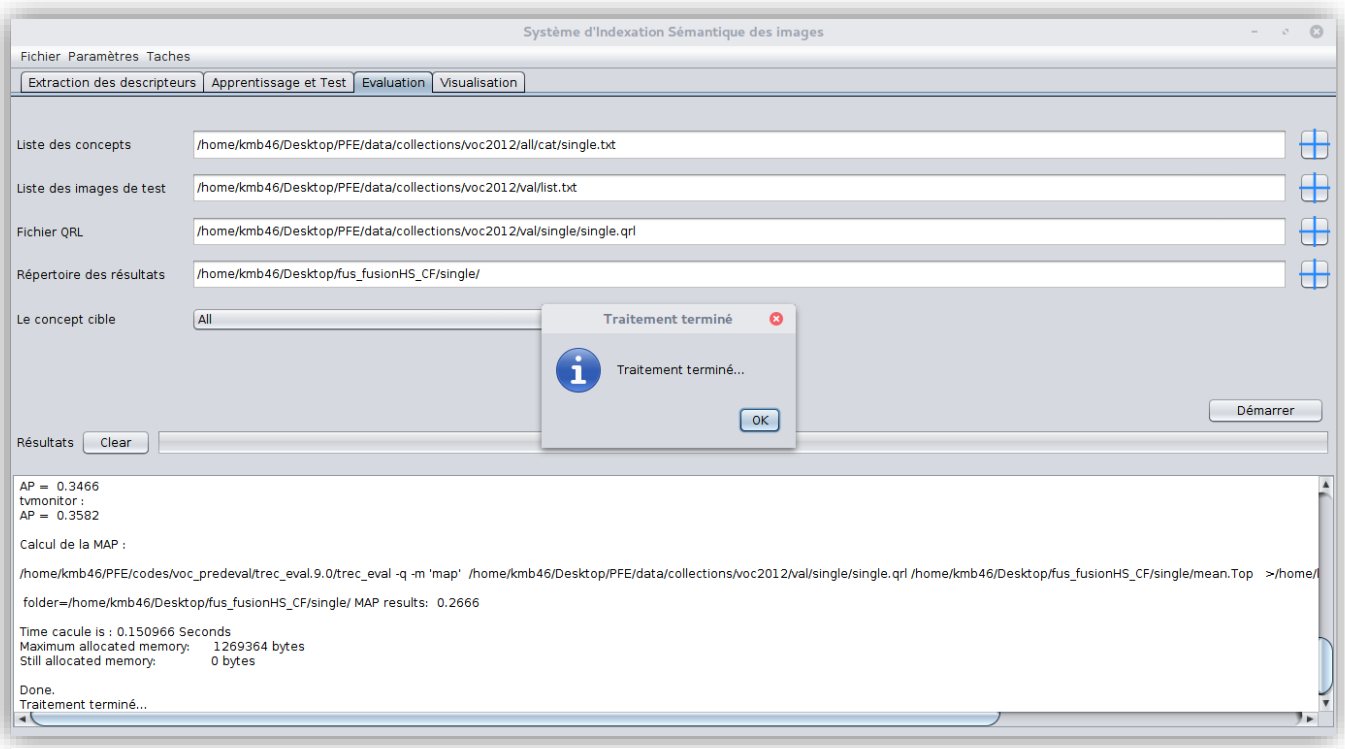

Figure 5.12 Interface graphique d'évaluation des résultats de détection de concepts

- La Figure 5.13 montre l'interface graphique de visualisation des résultats de détection de concepts.

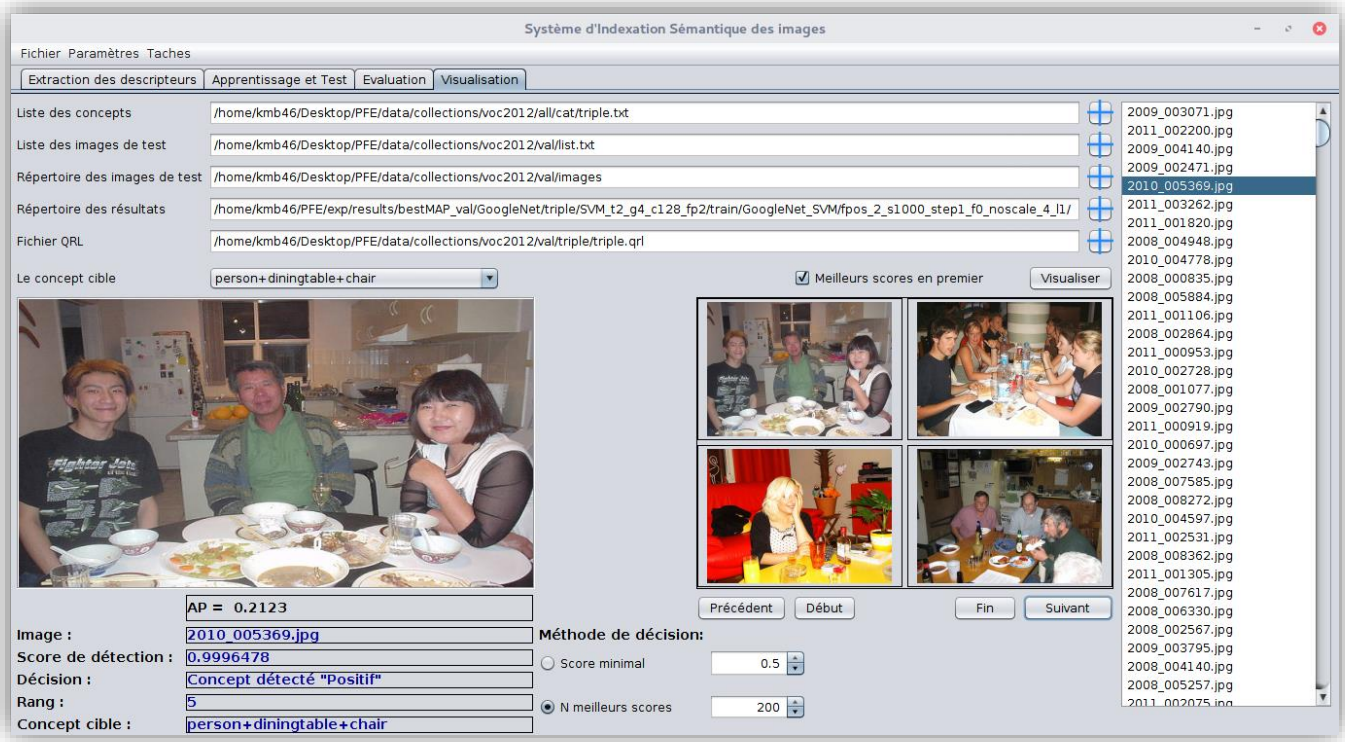

Figure 5.13 Interface graphique de visualisation des résultats de détection de concepts

5.4. Conclusion

A travers ce chapitre nous avons pu présenter les différentes étapes de conception de notre système de détection de concepts, en montrant les différents diagrammes UML utilisés. Nous avons présenté l'environnent de travail, et détaillé ensuite les différents langages et outils utilisés pour le développement de notre système. Nous avons terminé par une présentation des interfaces graphiques principales de fonctionnement du système. Dans ce qui suit nous allons présenter une conclusion générale.

Conclusion générale

La détection de concepts visuels dans les images est une tâche très importante pour concevoir des systèmes de recherche sémantique d'images. Bien que cette problématique soit très difficile, les performances des approches proposées dans l'état de l'art s'améliorent. Or, indexer des documents par des concepts singuliers ne suffit pas pour répondre à des requêtes complexes des utilisateurs qui comportent plusieurs sémantiques. Il est donc important de penser à la problématique de détection de plusieurs concepts simultanément (multi-concepts) dans les images afin d'aboutir à des résultats de recherche plus satisfaisants. Cette tâche a été très peu abordée dans l'état de l'art. D'autre part, en plus des descripteurs classiques (de bas niveau) utilisés dans les systèmes d'indexation des images, d'autres types de caractéristiques de haut niveau ont émergé et ont donné des résultats intéressants.

Dans le cadre de ce travail, nous avons réalisé un système de détection de concepts. Nous avons opté pour l'utilisation des descripteurs de haut niveau utilisant deux approches, la première fait l'extraction d'un descripteur sémantique utilisant la rétroaction conceptuelle, et la deuxième consiste à utiliser les résultats de l'apprentissage profond, par l'extraction d'un descripteur appris basé sur les résultats intermédiaires données par un modèle appris d'un DCCN.

Nous avons utilisé d'autres descripteurs qui sont de bas niveau, afin de faire une étude comparative les deux types de descripteurs (Descripteurs de bas vs haut niveau) dans le contexte de la détection des multi-concepts dans les images. Nous avons fait des expérimentations sur un corpus standard international « Pascal VOC 2012 » pour évaluer les performances des différents descripteurs.

Nous avons réalisé la majorité des composants qui peuvent être intégrés dans un système de détection de concepts. Après avoir évaluer les résultats de détection de paires et de triplets de concepts dans les images du corpus utilisé, nous avons déduit que l'utilisation de nos descripteurs sémantiques donne des performances significativement meilleures que celles de l'utilisation des descripteurs de bas niveau, et comparables à celles de l'état de l'art. D'autre part nous avons vu que l'utilisation des descripteurs appris donnent de très bonnes performances dans le contexte de la détection des multi-concepts, clairement meilleures que celles de l'utilisation des descripteurs de bas niveau, et parfois dépassent les performances de l'état de l'art dans la même tache.

Nous avons réalisé un système de détection générique, afin de rendre l'amélioration de ce système facile. Même si à travers nos expérimentations nous avons étudier plusieurs types de descripteurs de bas niveau, et plusieurs autres de haut niveau, l'utilisation d'autres descripteurs des deux types, peut donner encore d'autres résultats afin d'améliorer les performances de détection de multi-concepts dans les images.

Bibliographie

[1] A. Hamadi, Utilisation du contexte pour l'indexation sémantique des images et vidéos, Grenoble: Universite de Grenoble, 2014.

[2] A. Hamadi, P. Mulhem et G. Quenot, A comparative study for multiple visual concepts detection in images and videos, New York: Springer Science+Business Media, 2015.

[3] X. Li, C. G. M. Snoek, M. Worring et A. W. M. Smeulders, Harvesting Social Images for Bi-Concept Search, IEEE TRANSACTIONS ON MULTIMEDIA, 2012.

[4] M. Budnik, E.-L. Gutierrez-Gomez, B. Safadi et G. Quenot, Learned features versus engineered features for semantic video indexing, IEEE, 2015.

[5] B. Safadi et G. Quénot, Evaluations of multi-learners approaches for concepts indexing in video documents, Paris: RIAO, 2010.

[6] B. Safadi et G. Quenot, Re-ranking for Multimedia Indexing and Retrieval, Dublin, Ireland: European Conference on IR Research, 2011.

[7] D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 2004.

[8] B. Mateusz, G.-G. Efrain-Leonardo, S. Bahjat, P. Denis et Q. Georges, Learned features versus engineered features for multimedia indexing, Multimedia Tools and Applications, Springer Verlag, 2016.

[9] R. Ethan, R. Vincent, K. Kurt et B. Gary, ORB: an efficient alternative to SIFT or SURF, ICCV '11 Proceedings of the 2011 International Conference on Computer Vision , 2011.

[10] B. Herbert, E. Andreas, T. Tinne et V. G. Luc, Speeded-Up Robust Features (SURF), Computer Vision and Image Understanding, 2008.

[11] A. Hamadi, P. Mulhem et G. Quenot, Extended conceptual feedback for semantic multimedia indexing, Multimed Tools Appl, 2014.

[12] J. Sanchez, F. Perronnin, T. Mensink et J. Verbeek, Image Classification with the Fisher Vector: Theory and Practice, International Journal of Computer Vision, 2013.

[13] A. Krizhevsky, I. Sutskever et G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS, 2012.

[14] S. Christian, L. Wei, J. Yangqing, S. Pierre, R. Scott, A. Dragomir, E. Dumitru, V. Vincent et R. Andrew, Going Deeper with Convolutions, arXiv:1409.4842 [cs], 2014.

[15] H. Abdelkader, M. Philippe et Q. Georges, Conceptual feedback for semantic multimedia indexing, IEEE, 2013.