REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'enseignement supérieur et de la recherche scientifique

Faculté des Sciences

Département d'Informatique

Mémoire de Fin d'Etudes en vue de l'obtention du Diplôme de Master en Informatique

Options: Systèmes Informatiques (SI) – Réseaux et Systèmes Distribués (RSD)

Thème :

Recherche de l'Image par le Contenu Visuel: Une Approche par Apprentissage Profond (CNN)

Réalisé par :

Bourzam Nassim (SI)

Bougdah IssamEddine (RSD)

Encadré par :

Pr BOUCHEHAM Bachir

2021/2022

Remerciements

Nous remercions avant tout, Dieu de nous avoir donné la force morale et physique et nous a permis de terminer ce travail.

Nos vifs remerciements au Professeur BOUCHEHAM Bachir, notre encadreur, pour ses conseils et ses encouragements et sa patience avec nous.

Je remercie aussi les membres du jury qui ont accepté d'y participer et de juger ce travail.

Enfin, nous remercions les personnes les plus précieuses à nos yeux et dans notre vie, qui n'ont jamais hésité à nous apporter tout leur soutien, nos chers parents, nos frères et nos sœurs .Un grand merci à vous tous.

Merci.

Résumé

Résumé

Avec l'avènement des endoscopes médicaux, des satellites d'observation de la Terre et des téléphones personnels, la recherche d'images basée sur le contenu (CBIR) a attiré une attention considérable, déclenchée par sa large application, par exemple, l'analyse d'images médicales, la télédétection et la ré-identification de personnes. Cependant, la construction d'une extraction de caractéristiques efficace est toujours reconnue comme un problème difficile. Dans le deepn learning pour la recherche de l'image par le contenu, l'utilisation de ces techniques ont permis des progrès significatifs dans les domaines de traitement d'images pertinentes avec des résultats généralement pas toujours satisfaisants aux attentes des utilisateurs. Pour s'attaquer à cette problématique, nous proposons des solutions basées sur la VGG-16, cette méthode tentant de modéliser des données avec architectures complexes combinant différentes transformations non linéaires. Le VGG-16 est un réseau neuronal convolutif et également connu sous le nom de ConvNet, qui est une sorte de réseau neuronal artificiel utilisé comme algorithme pour mesurer le contraste de détection et de classification d'objets locaux d'une image. Il fut proposé par Karen Simonyan et Andrew Zisserman du Visual Géométrie Groupe Lab de l'Université d'Oxford en 2014, et a connu des succès notables dans divers domaines d'imagerie. Suite à celà, les créateurs de ce modèle ont évalué les réseaux et augmenté la profondeur en utilisant une architecture avec de très petits filtres de convolution (3 × 3), ce qui a montré une amélioration significative par rapport aux configurations de l'art antérieur. Ils ont poussé la profondeur à 16-19 couches de poids, ce qui en fait environ - 138 paramètres entraînables. VGG est basé sur la notion d'un réseau beaucoup plus profond avec des filtres plus petits. Aussi, VGG 16 est une architecture à 16 couches avec des couches convolutifs, une couche de regroupement, quelques couches convolutionnelles supplémentaires, une couche de regroupement, plusieurs couches de conversion supplémentaires, etc... Le résultat final est obtenu par VGG-16 sur la base Corel-10K. Nous comptons aussi étendre nos prospections de mesures de distance et mesures de similarité, montrent que le descripteur proposé est efficace, robuste et pratique en termes d'application CBIR.

Mot-Clé: CBIR, VGG-16, ImageNet, CNN, Deep learning, Computer vision.

Table des matières

Sommaire :

Sommaire

Chapitre 2 : Mesures de Similarité & Descripteurs d'Images

Liste des figures

Liste de Figures :

Liste Tableaux

Introduction Générale

De nos jours, les systèmes de vision sont de plus en plus utilisés dans La recherche visuelle par le contenu qui consiste à trouver dans une base de données, les images jugées les plus similaires à une requête donnée. Les performances d'un système de recherche d'images basé sur le contenu (Content Based Image Retrieval, CBIR) dépendent de manière cruciale de la représentation des caractéristiques et de la mesure de la similarité, qui ont été largement étudiées par les chercheurs en multimédia depuis des décennies. Bien qu'une variété de techniques ait été proposée, cela reste l'un des problèmes les plus difficiles de la recherche actuelle sur la recherche d'images basée sur le contenu (CBIR). D'autre part, le recours au CBIR en lieu et place de la recherche par mots clés (TBIR, Text Based Image Retrieval) est principalement justifié par le célèbre « fossé sémantique » existant obligatoirement entre les concepts véhiculés par les mots clés (utilisateur) et les résultats retournés par le système.

Suite à son succès, le CBIR s'est répandue dans plusieurs domaines : Médecine, Géologie, Astronomie, Tourisme, etc. Le CBIR consiste à extraire automatiquement les caractéristiques visuelles d'une image requête et de les comparer aux features extraites des images d'une base de données. Ce processus permet alors de déterminer parmi les images de la base, celle qui sont les plus similaires à l'image requête, du point de vue du système, bien sûr. Dans ce cadre, plusieurs techniques et méthodes ont été développées. Ces techniques sont regroupées essentiellement sur le contenu : Couleur, Texture et Forme, des objets contenus dans les images. Cependant, ces techniques restent applicables pour des collections d'images à sémantique réduite. Pour les datasets à sémantique très riche et diversifiée, d'autres techniques ont pris le relais.

Notamment, ces derniers temps ont été témoins d'avancées importantes de nouvelles techniques d'apprentissage automatique, en particulier, une technique importante connue sous le nom de « apprentissage en profondeur ». En dépit du succès rencontré par les méthodes de l'apprentissage en profondeur dans le domaine de la recherche d'images depuis 2012 [Kriz], les réseaux de neurones convolutifs profonds encodent une information locale des images relativement et peu adaptée à une telle problématique. Ce pendent les méthodes d'apprentissage en profondeur permettent à un système d'apprendre des fonctions complexes.

Dans le contexte de ce mémoire, nous nous focussons sur les notions générales de la recherche de l'image par le contenu visuel, particulièrement, à l'approche par apprentissage profond (CNN) qu'est un type d'intelligence artificielle, dérivée du machine Learning qui a été développé dans le but de créer des algorithmes capables d'apprendre et de s'améliorer de manière autonome.

Dans ce travail de Master, nous avons alors réalisé deux objectifs principaux : (a) Une étude théorique sur le CBIR, ses éléments essentiels, ses approches les plus significatives et particulièrement l'approche basée Deep Learning, (b) Côté contribution pratique, nous avons réalisé un CBIR basé

apprentissage profond (CNN) sur la base de l'architecture de l'apprentissage en profondeur pour la vision par ordinateur, la méthode VGG-16 CNN. Les tests furent effectués sur la base de données Corel-10K, bien répandue dans le contexte du CBIR. Cette base contient, au fat, 80 classes, mais, nous avons évalué notre système sur uniquement 10 classes, à cause des contraintes techniques. Les résultats obtenus montrent alors seulement la recevabilité de l'approche utilisée.

Le reste de ce mémoire est organisé comme suit :

Le premier chapitre : aborde d'une façon générale les Systèmes de recherche d'images par contenu (CBIR).

Le deuxième chapitre : Est consacré aux Mesures de Similarité & Descripteurs d'Images.

Le troisième chapitre : est dédié à l'apprentissage profond et vision par ordinateur.

Le dernier chapitre : Présente notre contribution pratique : Un système CBIR basé Deep Learning (CNNs).

Le mémoire se termine par une conclusion générale incluant des perspectives futures.

Chapitre 1 : Systèmes de recherche d'image par contenu (CBIR)

1. Introduction :

La définition « recherche d'images par le contenu » (« Content-Based Image Retrieval, CBIR, en Anglais) date aux travaux de Kato en 1992. Son système, ART MUSEUM, permet de retrouver des images d'art par couleurs et contours. Le terme s'est élargi par la suite à tout procédé permettant de rechercher des images selon des descripteurs utilisé, qui peuvent être de type « signal », comme la couleur et la forme, mais aussi symboliques. Comme le remarquent les auteurs d'un rapport important sur les systèmes de recherche par le contenu [Mera 2011], retrouver des images indexées manuellement par des mots clefs n'est pas de la recherche par le contenu au sens où le terme est généralement compris, même si ces mots clefs décrivent le contenu effectif de l'image.

Les applications des systèmes de recherche d'images existants (et donc les collections d'images) sont variées. Elles incluent des applications judiciaires : les services de police possèdent de grandes collections d'indices visuels (visages, empreintes) exploitables par des systèmes de recherche d'images.

Les applications militaires, bien que peu connues du grand public, sont sans doute les plus développées: reconnaissance d'engins ennemis via images radars, systèmes de guidage, identification de cibles via images satellites en sont des exemples connus. Le journalisme et la publicité sont également d'excellentes applications. Les agences de journalisme ou de publicité maintiennent en effet de grosses bases d'images afin d'illustrer leurs articles ou supports publicitaires. Cette communauté rassemble le plus grand nombre d'utilisateurs de recherche par le contenu (davantage pour les vidéos) mais l'aide apportée par ces systèmes n'est absolument pas à la hauteur des espoirs initiaux ([Mera2011]).D'autres applications incluent: le diagnostic médical, les systèmes d'information géographique, la gestion d'œuvres d'art, les moteurs de recherche d'images sur Internet et la gestion de photos personnelles.

Concevoir un système permettant d'assister des utilisateurs dans leurs tâches de recherche d'images pose des problèmes variés. Dans [Mera 2011] les difficultés suivantes sont identifiées :

1. Comprendre les besoins des utilisateurs d'images et leurs comportements : de quoi les utilisateurs ont-ils besoin ?

2. Identifier une manière « convenable » de décrire le contenu d'une image. C'est une tâche rendue difficile par la subjectivité intrinsèque aux images.

3. Extraire des « descripteurs » des images brutes.

4. Pouvoir stocker de manière compacte un grand nombre d'images.

5. Comparer requêtes et images stockées de manière à refléter les jugements de similarité humains.

6. Accéder efficacement aux images par leur contenu.

7. Fournir des interfaces utilisables.

Il convient donc d'abord de montrer quelles sont les approches existantes ainsi que leurs limitations. Nous commençons par rappeler les composants d'un système de recherche d'images par le contenu, puis nous présentons une taxonomie des systèmes selon leur niveau d'abstraction en donnant systématiquement des exemples.

2. Composants d'un CBIR :

Nous allons vous montrer ici les caractéristiques communes à la plupart des étapes : le traitement de la base d'images, les requêtes puis la mise en correspondance et la présentation des résultats. La Figure (Figure I.1) illustre l'ordonnancement de ces étapes:

Dans un premier temps (2), des descripteurs sont calculés à partir de chaque image de la collection (1), ils peuvent être de type signal ou/et symbolique (le vocabulaire d'indexation).Les données extraites (à présent représentatives du contenu de l'image du point de vue du système) constituent la base d'index (3). Les requêtes de l'utilisateur (4) sont alors transformées afin d'être comparables avec la base d'index (5) ; une mise en correspondance (6) entre la requête transformée et la base d'index permet ensuite de produire le résultat de la requête (7). Il se peut également que le système possède des composantes liées à la personnalisation, comme par exemple l'extraction, le stockage et l'utilisation d'un profil d'utilisateur.

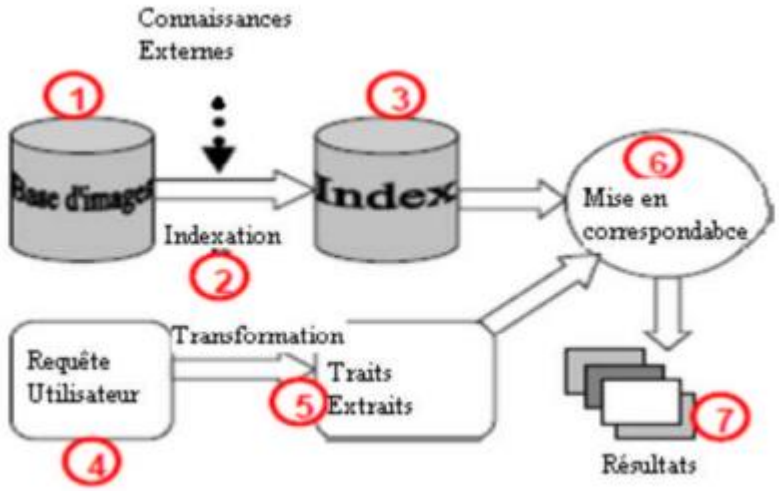

Figure I.1 : Principaux composants d'un Système de Recherche par le Contenu [Sito 1]

2.1 La base d'image :

Dans le cadre d'évaluer et de valider les systèmes de recherche de l'image par le contenu, les développeurs de ces systèmes ont recours à des bases standards utilisées par la communauté scientifique lors des tests et validations de leurs approches [Benl2013]. La plupart de ces bases d'images sont disponibles sur Internet librement. Dans le cas général, chaque base d'images possède des classes bien définies où chaque image n'appartient qu'à une seule classe. Dans certains cas une seule image peut appartenir à plusieurs classes. Pour permettre de faire cette identification, chaque image est alors accompagnée de ses métadonnées (données complémentaires qui servent à décrire le contenu de l'image).

2.1.1 Exemples de bases d'images existent :

Ils existent plusieurs bases d'images, on peut éclaircir les plus utilisé dans le domaine de recherche d'image par le contenu:

2.1.1.1La base de Wang :

La base d'images de Wang est un sous-ensemble de la base d'images Corel. Cette base d'images contient 1000 images naturelles en couleurs. Ces images ont été divisées en 10 classes, chaque classe contient 100 images. L'avantage de cette base est de pouvoir évaluer les résultats. Cette base d'images a été utilisée pour faire des expériences de classification. Un exemple de chaque classe peut être vu sur la figure (Figure I.2) [Wang et al., 2001] .Cette base d'images a été créée par le groupe du professeur Wang de l'université Pennsylvanie State et est disponible à l'adresse : http://wang.ist.psu.edu/.

Chaque image dans cette base d'images a une taille de 384 × 256 pixels ou 256 × 384 pixels.

Figure I.2 :10 classes de la base de Wang [Sito 2]

2.1.1.2 COIL (Columbia Object Image Library) :

Cette base d'images est très connue pour la reconnaissance des objets. Il y a deux bases d'images COIL : COIL-20 qui contient des images en niveaux de gris prises à partir de 20 objets différents et COIL-100 qui contient des images en couleurs prises à partir de 100 objets différents. Les deux bases d'images consistent en des images prises à partir des objets 3D avec des positions différentes. La base COIL-100 a 7200 images en couleurs (100 objets x 72 images/objet). Chaque image a une taille de 128×128 pixels. Tous les objets utilisés dans COIL-100 sont montrés dans la figure (Figure I.3) [Nayar et al., 1996].La base COIL-20 a 1440 images en niveaux de gris (20 objets x 72 images/objet). Chaque image a taille 128×128 pixels. Tous les objets utilisés dans COIL-20 sont dépeints sur la figure (Figure I.3) [Nayar et al., 1996]. Ces bases d'images sont disponibles à l'adresse:

http://www1.cs.columbia.edu/CAVE/research/softlib/

Figure I.3: Les objets utilisés dans COIL-100 [Sito 3]

2.1.1.3 Pollen :

La base de 347 images de grains de pollen (31 classes au total) de l'INRIA Sophia-Antipolis et provenant du projet Européen ASTHMA1. Quelques exemples de cette base d'images sont montrés dans la figure (Figure I.4).

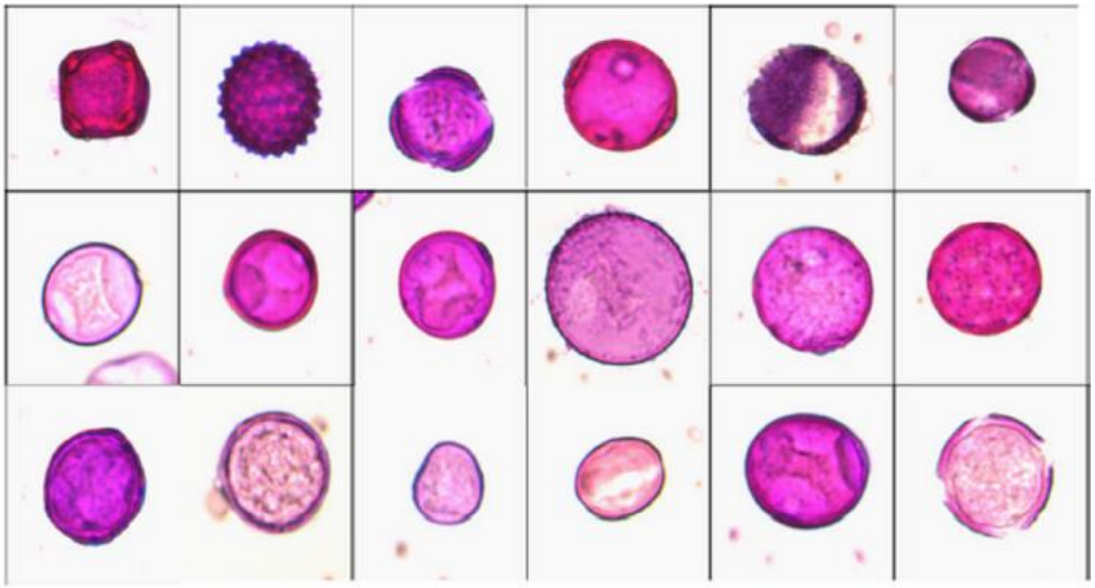

Figure I.4: Quelques images exemples de la base de Pollens [Sito 4]

2.1.1.4 CURET (Columbia Utrecht Reflectance and Texture Database) :

Cette base a plus de 14000 images de textures (70 textures x ~200 images/texture) des Universités Columbia et d'Utrecht (CUReT). Ces images sont disponibles à l'adresse :

http://www1.cs.columbia.edu/CAVE/curet.

Quelques exemples de cette base d'images sont montrés dans la figure Figure I.5.

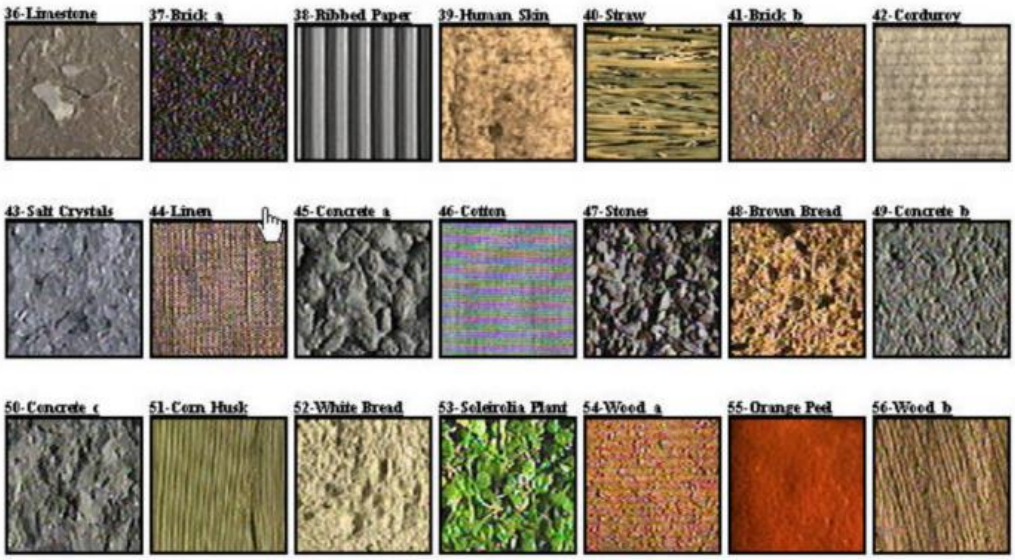

Figure I.5: Quelques images exemples dans la base de CUReT(CUReT) [Sito 5]

2.1.1.5 La base de FeiFei :

Cette base contient des images de 101 objets collectées par Fei-Fei Li, Marco Adreetto et Marc Aurolio Ranzato. Avec chaque objet, de 40 à 800 images ont été prises. Chaque image a une taille de 300×200 pixels. Ces images sont disponibles à l'adresse :

http://www.vision.caltech.edu/feifeili/Datasets.htm. La Figure I.6 montre quelques exemples de cette base :

Figure I.6 : Quelques images exemples dans la base de Fei-Fei [Sito 6]

2.2 L'indexation :

L'indexation est l'ensemble des processus aboutissant à la construction d'un index de l'image. Contrairement à d'autres types de données, comme le texte, il n'est pas utile d'utiliser les images directement dans un CBIR vue la taille des images actuelles. Il faut caractériser les images par des informations à la fois discriminantes et invariables à certains paramètres (comme la taille de l'image, l'angle de la prise de vue, etc.). L'indexation peut être fixe: les descripteurs calculés sont toujours les mêmes. L'indexation peut aussi être évolutive: les descripteurs s'adaptent à l'utilisateur ou au contexte dans le temps, ce qui permet de renforcer l'adéquation système/utilisateur. Cette caractérisation indispensable qui constitue l'une des taches importante dans le processus d'indexation, consiste à extraire automatiquement des caractéristiques à partir de l'image et les stocker dans un vecteur caractéristique visuelle (ou signature). L'indexation est donc le processus qui permet de construire un index de l'image [Bouk 2018]. En utilisant les techniques des bases de données, on peut stocker les caractéristiques de l'image et les récupérer rapidement et efficacement. Lorsqu'on travaille sur une base de taille assez importante, la gestion des index que ce soit pour le stockage ou l'accès devient nécessaire. Il y a deux manière pour stocker l'index : séquentielle ou hiérarchique [Bouk 2018]. Lorsque le nombre d'image n'est pas assez grand, on opte pour la méthode séquentielle, que ce soit en mémoire ou dans un fichier. Cependant, lorsque le nombre d'image augmente, il devient plus pratique d'organiser les index de manière hiérarchique [Bouk 2018], sous forme d'arbres ou de tables afin d'accélérer l'accès à l'information.

2.3 La gestion des index :

Elle concerne la manière dont sont gérés les index des images : stockage et accès. La gestion des index, anecdotique pour une collection de taille modeste, devient une préoccupation essentielle lorsque l'on travaille sur une base de taille conséquente. La manière la plus basique de stocker les index est la liste séquentielle, que ce soit en mémoire ou dans un fichier. Cependant, lorsque le nombre d'images augmente, le temps d'accès à une image augmente linéairement et il est souvent nécessaire d'organiser les index de manière hiérarchique, sous forme d'arbres (organisés selon les descripteurs), ou de tables de « hash-code » par exemple, afin d'accélérer l'accès à l'information.

2.4 Les requête :

Le type de requête proposé découle de choix fait en amont, au niveau de l'indexation. Dans des systèmes où seuls des descripteurs de bas niveau sont extraits, les requêtes ne peuvent être que de bas niveau : requête par « image exemple », par croquis ou par manipulation directe des traits de bas niveau. Dans ces systèmes, des descripteurs sont extraits à partir de la requête (une image, un croquis…) et sont comparés aux descripteurs calculés à partir des images de la base (les index des images).

A l'opposé, dans des systèmes proposant plus d'abstraction, les requêtes peuvent être sémantiques (textuelles par exemple). Par exemple [Mera et Mah 2011], les images sont indexées par des « catégories sémantiques visuelles », ce qui permet à un utilisateur de formuler des requêtes sémantiques (« Je veux des images prises à l'extérieur. »).

2.4.1 Les types des requêtes :

La requête adressée au système de recherche doit permettre à ce dernier de retrouver les images désirées par l'utilisateur. Suivant les besoins de l'utilisateur et le type de base de données, plusieurs types de requêtes sont possibles, les plus populaires sont : la requête par esquisse (croquis), la requête par exemple et la combinaison de celles-ci afin d'accéder à un niveau d'abstraction supérieur [Benl2013].

2.4.1.1 Requête par l'exemple :

Pour représenter ses besoins, l'utilisateur fournit une image ou une partie d'une image qu'il considère similaire aux images qu'il souhaite rechercher. Cette image est appelée image exemple ou requête. L'image exemple est soit fournie par l'utilisateur, soit choisie par ce dernier dans la base d'images utilisée. Cette technique est simple et ne nécessite pas de connaissances approfondies pour manipuler le système. La plupart des systèmes de recherche de l'image par le contenu utilisent cette technique. La simplicité de son concept garantit un accès à grande échelle aux images indexées. C'est l'un des buts essentiels de la recherche de l'image par le contenu [Benl2013].

2.4.1.2 Requête par crayonnage (Sketch) :

Dans ce cas, le système fournit à l'utilisateur des outils lui permettant de constituer une esquisse (en Anglais : Sketch) qui correspond à ses besoins. L'esquisse fournie sera utilisée comme exemple pour la recherche. L'esquisse peut être une ébauche de la forme ou contour d'une image entière ou une ébauche des couleurs ou textures des régions d'une image. L'utilisateur choisira, en fonction de la base d'images utilisée et de ses besoins et préférences la représentation adéquate. Certains systèmes offrent même la possibilité de créer des esquisses tridimensionnelles [Benl2013]. La technique de la requête par l'esquisse présente un inconvénient majeur : il est parfois très difficile pour l'utilisateur de bien formuler sa requête, le fait d'avoir un large panel d'outils de dessin ne garantit pas forcément un bon résultat.

2.4.1.3 Requête par caractéristique :

L'utilisateur indique la ou les caractéristiques qu'il veut utiliser pour trouver les images similaires, par exemple trouver les images contenant 25% de rouge et 30% de jaune. Ces caractéristiques sont répertoriées dans un vocabulaire compilé en outils de traitement.

2.4.1.4 Requête exemple et texte :

Cette méthode consiste à renforcer l'image requête en lui associant du texte afin d'accéder à un niveau d'abstraction supérieur. Les images sont organisées et indexés en groupes de pertinence. Chaque groupe définit un type ou un domaine dont l'image relève, par exemple végétation ou animale, etc. Cette approche permet une sémantique accrue mais elle est plus biaisée voire moins générale que les précédentes en ce que la désignation des groupes de pertinence est une tâche à fort caractère subjectif ou expert.

2.5 Analyse de la requête :

Après que l'utilisateur introduit sa requête, le système procède à l'analyser et la transformer pour la rendre comparable avec l'index de la base d'image. La transformation dans ce cas consiste à extraire des descripteurs de même type que ceux extraits de la base d'image lors de l'indexation [Bouk2018].

2.6 Mise en correspondance requête / base :

Il s'agit d'estimer dans quelle mesure une image (son index) satisfait une requête donnée. Dans le contexte de la recherche d'images, cela se ramène souvent à calculer la similarité entre les caractéristiques extraites de la requête et les caractéristiques de chaque image dans la base. Cela aboutit généralement à une valeur de correspondance qui caractérise la pertinence (du point de vue du système) d'une image par rapport à la requête. Cette mise en correspondance peut être simple (comparaison d'histogrammes) ou complexe (comme par exemple [J. R. Smith1996], avec une mise en correspondance qui tient compte de l'arrangement spatial des régions).

La phase de mise en correspondance peut également inclure une pondération des descripteurs (comme dans chaque descripteur est pondéré par rapport à son pouvoir discriminant dans la base). Pondérer les descripteurs permet d'éliminer une partie du bruit dans la mesure où les descripteurs les moins pertinents voient leur influence diminuer dans l'évaluation de la similarité requête/image

La mise en correspondance peut également inclure un bouclage de pertinence. Le but est également d'éliminer le bruit (augmenter la précision) en tentant de converger vers une précision maximale.

2.7 La présentation des résultats :

Dans la grande majorité des systèmes disponibles [Mera et Mah 2011], le résultat d'une requête est présenté sous la forme d'une liste d'images (réduites à des vignettes) ordonnées par pertinence décroissante. Parfois cette présentation prend d'autres formes, comme par exemple l'œil de poisson (FishEye View) [Mera et Mah 2011]. L'avantage des images par rapport aux documents textuels est qu'il est possible de visionner d'un coup d'œil l'intégralité du document, ce qui permet de visualiser un grand nombre de résultats et de les comparer plus rapidement. Comme indiqué plus haut, la présentation des résultats est souvent couplée avec une possibilité d'interaction, qui permet par exemple de raffiner une requête en indiquant au système les résultats pertinents et ceux qui ne le sont pas (bouclage de pertinence), et de permettre ainsi une reformulation automatique de la requête.

3. Représentation des images dans un CBIR :

Dans la majeure partie des systèmes existant, les images sont représentées avec des descripteurs de bas niveau, i.e., en termes de couleur, texture, formes (voir le chapitre suivant) ou par des descripteurs de haut niveaux.

3.1 Descripteurs de bas niveau :

Pour pouvoir utiliser ces informations, il faut réduire de manière drastique cette quantité. C'est là où vien le rôle de la perception : extraire d'une quantité énorme d'information utilisable est cela consiste essentiellement à trouver des régularités dans les données. Les régularités sont intéressantes car elles permettent de coder, représenter une information brute, de manière concise (particulièrement lorsqu'on ignore les détails). Prenons l'exemple de notre système visuel. Il existe toutes sortes de régularités visuelles et nous possédons un certain nombre de détecteurs pour ces régularités. Il se trouve que les régularités que nous sommes capables de détecter nous permettent d'avoir suffisamment d'information pour identifier la plupart des objets physiques. Dans ce chapitre, nous nous intéressons aux techniques utilisées en informatique pour reproduire notre processus de perception. Nous appellerons ces régularités dans l'information visuelle des descripteurs de bas niveau [Mera et Mah 2011].

4. Mesures pour évaluer un système :

Avant l'exécution d'un système de recherche d'informations, une évaluation qui permet de mesurer la performance de ce système est nécessaire. Les mesures les plus courantes pour évaluer un système sont le temps de réponse et l'espace utilisé. Plus le temps de réponse est court, plus l'espace utilisé est petit, et plus le système est considéré bon. Mais avec des systèmes qui ont été faits pour la recherche d'informations, en plus de ces deux mesures, on s'intéresse à d'autres mesures. Dans le système de recherche d'informations, l'utilisateur s'intéresse aux réponses pertinentes du système. Donc les systèmes de recherche d'informations exigent l'évaluation de la précision de la réponse. Ce type d'évaluation est considéré comme l'évaluation des performances de recherche. Le système d'indexation et de recherche d'images est un système de recherche d'informations. Dans les systèmes de recherche d'images, les auteurs ont souvent utilisé les mesures d'évaluation pour évaluer des systèmes de recherche d'informations.

Dans cette section, nous allons décrire les deux mesures les plus courantes: le rappel et la précision. Ces mesures sont reliées entre elles. Donc on décrit souvent cette relation par une courbe de rappel et précision. Ensuite nous présentons d'autres mesures que l'on utilise aussi pour évaluer des systèmes de recherche d'informations.

4.1. Rappel et précision (en anglais : Recall and Precision) :

2 valeurs pour évaluer le système de recherche par contenu :

4.1.1. Le rappel:

Le rappel est le rapport entre le nombre d'images pertinentes dans l'ensemble des images trouvées et le nombre d'images pertinentes dans la base d'images.

$$Rappel = \frac{|Ra|}{|R|} \qquad (4.1.1)$$

4.1.2. La précision :

La précision est le rapport entre le nombre d'images pertinentes dans l'ensemble des images trouvées et le nombre d'images trouvées.

$$\text{Précision} = \frac{|Ra|}{|A|} \qquad (4.1.2)$$

Où :

_ R : l'ensemble d'images pertinentes dans la base d'images utilisée pour évaluer.

_ |R| : le nombre d'images pertinentes dans la base d'images.

_ A : l'ensemble des réponses.

_ |A| : le nombre d'images dans l'ensemble des réponses.

_ |Ra| : le nombre d'images pertinentes dans l'ensemble des réponses.

Des définitions sont montrées dans la Figure I.7:

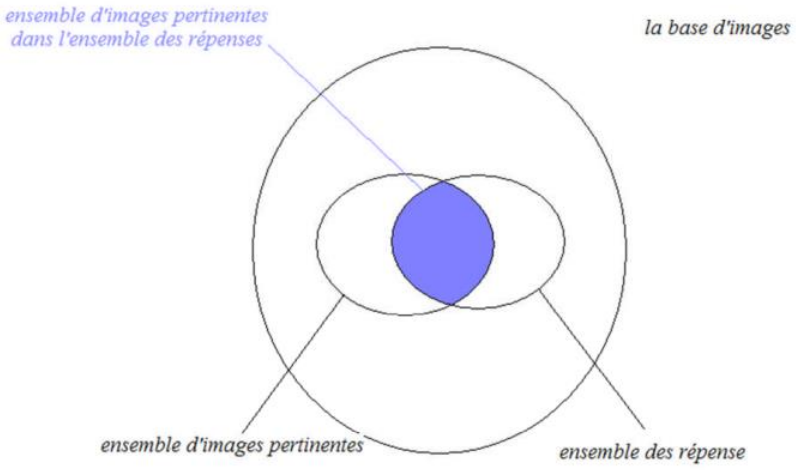

Figure I.7: Le rappel et la précision pour une requête [Sito 7]

Dans les systèmes de recherche d'informations, afin de définir si une information est pertinente ou non, on a besoin d'experts dans le domaine. Dans les systèmes de recherche d'images, une image est pertinente pour une requête si les deux images sont dans la même classe. C'est pourquoi dans l'étape de préparation de la base d'images pour évaluer, on doit faire des annotations. L'annotation est un processus qui permet aux utilisateurs de choisir des mots clés correspondants à chaque image. Après l'annotation, on va classifier les images en classes appropriées. Si des images ne contiennent pas beaucoup d'objets, c'est facile de les classifier dans ces classes. Mais si les images contiennent beaucoup d'objets, la tâche de classification devient de plus en plus difficile. Dans ce cas-là, chaque image appartient à plusieurs classes.

5. Conclusion :

Dans ce chapitre nous avons essayé de décrire les concepts de base pour la construction d'un système de recherche d'images par le contenu. Nous avons abordé dans un premier lieu les différents composants d'un CBIR. Par la suite nous avons parlé Représentation des images dans un CBIR, les

mesures pour évaluer un CBIR, difficulté de l'indexation des images. À la fin de ce chapitre, nous avons présenté les domaines d'application de la recherche d'images par le contenu. Comme nous pouvons le constater, la recherche d'images par le contenu, s'est imposée dans tous les domaines de notre vie quotidienne. Par conséquent elle attire beaucoup d'attention et devient un axe de recherche très actif avec tous ses aspects. Dans le prochain chapitre, nous présenterons les différents descripteurs extraient à partir d'une image (de couleurs, de texture et de forme) et les mesures de similarité entre ces descripteurs.

Chapitre 2 : Mesures de Similarité & Descripteurs d'Images

1. Introduction :

Aujourd'hui avec le développement des systèmes multimédias et le recul de l'écrit, nous utilisons de plus en plus le contenu visuel comme support de communication dans différents domaines. En effet l'image et la vidéo numérique sont partie intégrante de tels systèmes par la densité et la richesse de leur contenu. La même image peut présenter plusieurs significations à différents niveaux : analyse, description, reconnaissance et interprétation.

La recherche d'information couvre le traitement de documents numériques impliquant la structure, l'analyse, le stockage et l'organisation des données. Dans le passé, le terme recherche d'information était lié au concept de l'information textuelle. Actuellement « RI »est associé à tout type d'information, textuelle, visuelle ou autre. Cependant dû aux limitations des méthodes textuelles, le développement des méthodes basées sur le contenu visuel est devenu primordial. Ceci explique l'activité de recherche intense consacrée au système CBIR ces dernières années. Le « RIC » est souvent confronté au problème de pertinence de la recherche, et au temps de recherche.

L'objectif de n'importe quel système CBIR est de satisfaire la requête d'un utilisateur par la pertinence des résultats. Comme l'accès à un document via sa pure sémantique est impossible, les systèmes CBIR traditionnels s'appuient sur un paradigme de représentation de bas niveau du contenu de l'image, par la couleur, la texture, la forme, etc.…, et d'autres par une combinaison de celles-ci. La recherche d'images se fait ainsi par comparaison des descripteurs.

L'analyse et la représentation du contenu des données sources mises sous forme de vecteur caractéristique. L'information obtenue dans cette étape est une sorte de résumé des images de la base (segmentation en régions, couleur, texture, relations spatiales,…). La transformation est généralement gourmande en temps de calcul.

Dans la suite de ce chapitre, nous présentons les différents attributs utilisés dans les systèmes de recherche d'image par contenu et ensuite les mesures de similarité entre les images après la définition de leurs descripteurs.

2. Descripteurs d'image :

2.1. Descripteurs de couleur :

2.1.1. L'espace de couleur :

Chaque pixel d'image peut être représenté comme un point dans un espace 3D. Les espaces les plus communément utilisés dans les CBIR sont: l'espace RGB, Cie L*a*b*, CIE L*u*v*, HSV (ou HSL, HSB).

Typiquement les images sont codés sur trois canaux contre un seul pour les monochromes. Il semble que son efficacité soit liée au fait que l'être humain peut distinguer des milliers de couleurs et seulement 24 niveaux de gris.

Plusieurs études ont été réalisées sur l'identification d'espaces colorimétriques plus discriminants [Burger et Burge, 2009]. Par exemple la projection de l'image dans l'espace HSV permet de séparer les informations relatives à la teinte, la saturation et l'intensité [Burger et Burge, 2009], [Datta et al., 2008]. Il a été démontré que la teinte est mieux invariante aux conditions d'éclairage et de prise de vue.

D'autres espaces également fréquents dans le domaine revendiquent d'être perpétuellement uniformes et indépendant de l'intensité telle que CIE, XYZ, et CIE-LUV. Là encore ce sont des modèles de représentations et il n'existe pas un espace de couleur idéal. On trouve une comparaison entre les espaces de couleurs ainsi que leurs caractéristiques et une analyse avantages /inconvénients dans [Datta et al., 2008].

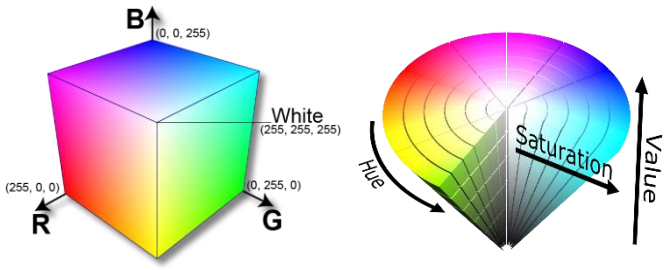

Figure II. 1: Espace de couleur RGB et HSV **[Sito 8]**

- **Système RGB :**

Le système le plus couramment utilisé est le système RGB (Red-Green-Blue), qui est le système des trois couleurs fondamentales. Il associe à chaque couleur trois composantes (ou canaux), qui correspondent aux intensités respectives de trois couleurs primaires de la synthèse additive. Le blanc correspond à la valeur maximale pour chaque canal, tandis que le noir correspond aux trois composantes nulles. En pratique, les valeurs de chaque canal sont des entiers compris entre 0 et NR pour le canal rouge, entre 0 et NG pour le canal vert et entre 0 et NB pour le canal bleu. Ainsi, chaque couleur appartient à un parallélépipède. Le codage le plus couramment utilisé consiste à prendre NR = NG= NB=255, ce qui permet de stocker chaque composante sur un octet. C'est ce codage qui est principalement utilisé.

- **Système normalisé (r, g, b):**

Un des problèmes du système RGB est que trois canaux ne représentent pas seulement la couleur, mais aussi la luminosité. Le Système (r, g, b) permet de supprimer en partie en normalisant les trois composantes R, G, et B. On obtient alors trois canaux r, g et b, ce dernier étant facultatif car il peut être obtenu par combinaison linéaire des deux autres.

Ainsi, on travaille en générale uniquement sur les canaux r et g, que l'on appelle système normalisé r, g, b.

Le passage du système RGB au système normalisé (r, g, b) s'effectue en divisant chaque composante par la somme des trois :

$$\begin{cases} r = \frac{R}{R+G+B} \\ g = \frac{G}{R+G+B} \\ b = \frac{B}{R+G+B} \end{cases} \qquad (2.1.1.1)$$

Ainsi, la somme des trois nouvelles composantes est égale à 1. Chaque couleur normalisée est un réel appartenant à l'intervalle [0,1].

- **Système XYZ:**

La Cie (Commission Internationale de l'éclairage) est une organisation internationale chargée d'établir des normes et des recommandations reconnues par tous les pays, afin de pouvoir quantifier la couleur : c'est la base de la colorimétrie, science de la mesure de la couleur. Le système XYZ a été établi par la CIE afin de pallier à certain inconvénients du système RGB. Ce système correspond à un changement de couleurs primaires et s'obtient simplement à partir du système RGB à l'aide d'une matrice de passage. Les coefficients de cette matrice sont déterminés par rapport à un blanc de référence que l'on appelle illuminant. Celui que nous utilisons est appelé illuminant standard D65. Il existe ainsi plusieurs codage XYZ, qui dépendent de l'illuminant choisi.

Ce système de couleurs ne nous intéresse pas particulièrement, mais il constitue une étape intermédiaire pour passer au système L*u*v.

Le passage du système RGB que nous utilisons (tous les canaux sont codés par des entiers entre 0 et 255) au système XYZ se fait par le calcul matriciel suivant :

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.430574 & 0.341550 & 0.178325 \\ 0.222015 & 0.706655 & 0.071330 \\ 0.020183 & 0.129553 & 0.939180 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \qquad (2.1.1.2)$$

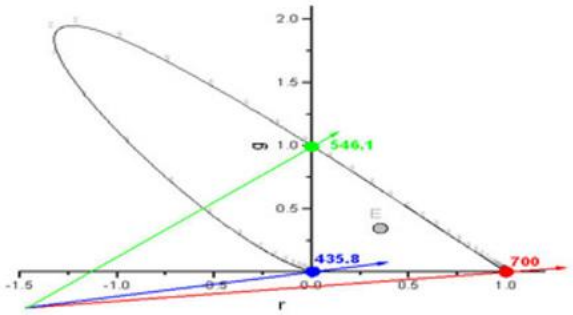

Figure II.2 : Espace XYZ/RGB. **[Sito 9]**

- **Système HSV:**

Le système (Hue –Saturation –Value) est défini par un cylindre qui représente la teinte, la saturation et la valeur d'une couleur. La teinte H est représentée par un angle entre 0 et 360° : elle indique la famille de couleur (rouge, jaune, vert, bleu, etc.). La saturation S donne une information sur la pureté de la couleur. La valeur V correspond à l'intensité lumineuse. Elle indique si la couleur est claire ou sombre. L'algorithme de transformation du système RGB vers HSV est donné par [Heus] :

(1) Max = max(R, G, B)

(2) Min = min(R, G, B)

(3) V = max

V=max(R, G, B)

S= (V-min(R, G, B))*255/V if V !=0 , otherwise

(G-B)*60/S, if V=R

H=180+ (B-R)*60/S, if V=G

240+(R-G)*60/S, if V=B

If H < 0 then H = H+360

(4)Si Max = 0

S=0

Sinon

S=(Max–Min)/Max

(5) Si Max=Min

H=0

Sinon

Si Max=R Et G≥B

H=60(G-B)/(Max-Min)

Sinon Si Max=R Et G<B

H=360+60(G-B)/(Max–Min)

Sinon Si G=Max

H=60(2+(B-R)/(Max–Min))

Sinon

H = 60(4 + (R-G)/(Max-Min)

Pour obtenir des valeurs entières entre 0 et 255, la bibliothèque Open CV propose les conversions suivantes :

$$V = \max(R, G, B) \qquad (2.1.1.3)$$

$$S = \frac{255(V - \min(R,G,B))}{V} \quad si\ V \neq 0, \quad 0\ sinon \qquad (2.1.1.4)$$

$$H = \begin{cases} \frac{60(G-B)}{S} & si\ V = R \\ 180 + \frac{60\,(B-R)}{S} & si\ V = G \\ 240 + \frac{60(R-G)}{S} & si\ V = B \end{cases} \qquad (2.1.1.5)$$

Si H<0 alors H=H+360

H = H/2 (pour obtenir une valeur qui tienne sur un octet).

2.1.2. Histogrammes :

Une méthode utilisée pour la couleur est l'intersection d'histogrammes [Y.Run, 1999]. Les histogrammes sont faciles et rapides à calculer, et robustes à la rotation et à la translation. Cependant l'utilisation d'histogrammes pour l'indexation et la recherche d'images pose quatre problèmes. Premièrement, ils sont de grandes tailles, donc par conséquent il est difficile de créer une indexation rapide et efficace en les utilisant tels qu'ils sont. Deuxièmement, ils ne possèdent pas d'informations spatiales sur les positions des couleurs. Troisièmement, ils sont sensibles à de petits changements de luminosité, ce qui est problématique pour comparer des images similaires, mais acquises dans des conditions différentes. Et quatrièmement, ils sont inutilisables pour la comparaison partielle des images (objet particulier dans une image), puisque calculés globalement sur toute l'image.

2.1.2. Les moments de couleur :

Les moments de couleur ont été utilisés dans plusieurs systèmes de recherche d'images par le contenu tel que QBIC, mathématiquement les trois premiers moments sont définis par :

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} f_{ij} \qquad (2.1.2)$$

$$\sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N}(f_{ij} - \mu_i)^2\right)^{1/2} \qquad (2.1.3)$$

$$s_i = \left(\frac{1}{N}\sum_{j=1}^{N}(f_{ij} - \mu_i)^3\right)^{1/3} \qquad (2.1.4)$$

Où f_{ij} est la valeur de la i ème composante chromatique du pixel j, et N le nombre de pixels de l'image. Les moments de couleur est une représentation compacte comparée aux autres descripteurs de couleur. Car seulement 9 valeurs (3 pour chaque composante chromatique) sont utilisées pour représenter le contenu d'une image. Pour cette raison ils peuvent diminuer le pouvoir de discrimination (description). [Mesk, 2009]

2.1.3. Cohérence spatiale :

Ce descripteur a pour but de combler en partie, l'absence d'information spatio-colorimétrique de l'image dans le descripteur précédent. La cohérence spatiale est calculée pour chaque classe de couleur identifiée. Tout d'abord un histogramme de connexité est calculé :

$$H_I(c) = \sum_{i=0}^{X=I}\sum_{j=0}^{X=I}\delta(I(i,j),c)\alpha(i,j) \qquad (2.1.5)$$

I est l'image segmentée de taille (X, Y), c'est la couleur du pixel (i, j), est le symbole de Kronecker et définie par

$$\alpha(i,j) = \begin{cases} 1 \; si \;\; \forall k,k' \in (-W,W) I(i+k,j+k') + I(i,j) \\ 0 \quad sinon \end{cases} \qquad (2.1.6)$$

La fenêtre (2W+1)*(2W+1) représentant le degré de compacité souhaité. La cohérence spatiale est alors donnée par le rapport :

$$SCR(c) = \frac{H_I(c)}{H(c)} \qquad (2.1.7)$$

Où H représente l'histogramme couleur et donc SRC(c) Є [0,1]. Une faible valeur de SRC(c) indiquera que la couleur c est dispersée dans l'image, tandis que pour une couleur dominante homogène SRC(c) sera proche de 1. [Houa, 2010]

2.1.4. Couleurs dominantes :

L'utilisation d'histogrammes pour représenter la distribution de couleur présente quelques inconvénients. Du point de vue de l'espace mémoire, les histogrammes à plusieurs dimensions sont ((creux)), c'est-à-dire que la majorité des cellules ne comptent aucun pixel. Une grande partie de l'espace mémoire est utilisée inutilement. De plus, toutes les classes ont la même taille, alors qu'il serait plus intéressant d'avoir des classes plus petites dans les régions contenant des couleurs très fréquentes, et de grandes classes pour les couleurs moins répandues. Du point de vue des mesures de similarité employées, les mesures traditionnelles effectuent uniquement une comparaison cellule à cellule. Même si les histogrammes sont ordonnés, le voisinage des cellules n'est pas pris en compte quand elles ont des valeurs différentes

Les signatures par couleurs dominantes, proposées dans [Lai], permettent de résoudre ces différents problèmes. La signature $s = \{s_i = (m_i, w_i)\}$ est un ensemble de nuages de points. Chaque nuage est représenté par son mode m_i (le mode d'un nuage de point correspond à un maximum local de sa densité de probabilité), et le nombre w_i de pixels qui appartiennent au nuage.

Contrairement aux histogrammes, ces signatures ne stockent que les couleurs qui appartiennent à l'image, elles ne stockent pas les cellules vides.

La norme MPEG-7 définit un descripteur pour les couleurs dominantes, appelée DCD (Dominant Color Descriptor). D'après [Wu], ce descripteur est défini par :

$$F = \{(c_i, p_i, v_i), s\} \qquad i = 1,2, \dots, N$$

Où N est le nombre de couleurs dominantes (inférieur ou égal à huit). Le terme C_i est un vecteur qui représente la $i^{ème}$ couleur dominante, attribuée à un pourcentage p_i de pixels dans l'image (ou dans la région de l'image). Le terme v_i représente un paramètre optionnel, la variance des couleurs des pixels qui sont associés à la couleur dominant C_i. Le terme s est une valeur qui représente l'homogénéité spatiale des couleurs dominantes dans l'image (en termes de connexité de régions). Le nombre d'occurrences de couleurs dominantes est variable selon les spécifications MPEG-7. Etant donné que le calcul de distance est assez complexe et pour des raisons de normalisation des calculs dans notre implémentation qui vise justement à réduire le temps de calcul, ce descripteur est peu intéressant dans un contexte d'indexation temps réel.

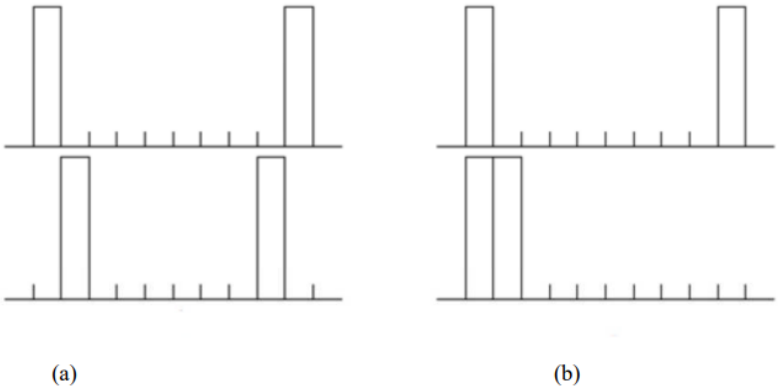

Figure II.3 : Comparaison entre histogrammes **[Sito 10]**

Avec une comparaison cellule à cellule, les deux histogrammes dans la situation (a) ont une intersection nulle, alors qu'ils sont très proches. Dans la situation (b), les deux histogrammes ont une intersection de 50% alors qu'ils ne sont pas visuellement plus proches que dans le cas (a), lorsque les cellules sont ordonnées par proximité de couleurs.

La mesure de similarité la plus utilisée pour comparer ces signatures est l'Earth Mover Distance, ou EMD. La distance entre deux distributions S1 et S2 est donnée par le coût minimum de travail nécessaire pour transformerS1 en S2 L'EMD peut être définie comme la solution d'un problème de transport qui peut être résolu par une optimisation linéaire.

Si on note d_{ij} la distance entre le modela m_j de s1 et le mode m_j de s2, et f_{ij} le point optimal entre les deux distributions, la solution de ce problème est donnée par :

$$EMD(s_1, s_2) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij}} \qquad (2.1.4)$$

Où m et n représentent respectivement le nombre de classes dans S1 et S2. Pour avoir plus de détails sur les problèmes d'optimisation de l'EMD, le lecteur peut se référer à [Lai].

Il existe plusieurs approches pour trouver les différents nuages de points. La méthode la plus simple consiste, à partir d'un histogramme, à prendre toutes les classes qui ont un effectif non nul.

Cette approche a l'inconvénient de fournir un nombre assez élevée de classes. Une autre approche consiste à ne garder que les couleurs dominantes, et à remplacer chaque couleur de l'image par la couleur conservée la plus proche. Il existe plusieurs algorithmes pour détecter les couleurs dominantes : citons notamment la segmentation d'image, qui consiste à regrouper tous les pixels ayant une couleur proche,

ou encore l'extracteur de couleurs dominantes proposée pour le descripteur de MPEG-7 par [Deng]. Ce dernier effectue une classification des couleurs de l'image dans un espace perceptuelle ment uniforme (habituellement le système L*u*v*), à l'aide de l'algorithme de Lloyd généralisé.

Même si l'EMD présente de bons résultats en comparaison d'autres mesures de similarité [Agro], son utilisation est limitée par la complexité de calcul. Même s'il existe des implémentations optimisées assez rapides, l'exécution demeure toujours plus lente qu'avec les distances traditionnelles appliquées sur les histogrammes. [Houa, 2010]

Conclusion :

L'information relative aux couleurs est particulièrement importante dans la caractérisation d'une image. Plusieurs études ont été menées pour trouver un critère de choix des descripteurs de couleurs pour l'indexation des images, mais aucune n'a abouti. Ceci peut s'expliquer par le manque de subjectivité de cette information, les descripteurs couleur ne suffisent pas à indexer efficacement une image, ni à la chercher.

2.2. Descripteurs des textures :

Au même titre que la couleur, la texture est une caractéristique fondamentale des images car elle concerne un élément important de la vision humaine. De nombreuses recherches ont été menées à la fois dans les domaines de l'analyse et de la synthèse de texture.

L'étude de la texture des objets d'une image peut avoir des objectifs très divers : obtenir des informations sur la nature d'un l'objet, segmenter l'image en régions homogènes, identifier la texture afin de la réduire à un ensemble de paramètres (compression d'images), recherche d'image par contenu, etc.

D'après [Bimb], une définition formelle de la texture est quasiment impossible.

D'une manière générale, la texture se traduit par un arrangement spatial des pixels que l'intensité ou la couleur seules ne suffisent pas à décrire. Elles peuvent consister en un placement structuré d'éléments mais peuvent aussi n'avoir aucun élément répétitif.

De nombreuses définitions ont été proposées, mais aucune ne convient parfaitement aux différents types de textures rencontrées. Dans une définition couramment citée [Fabio], la texture est présentée comme une structure disposant de certaines propriétés spatiales homogènes et invariantes par translation. Cette définition stipule que la texture donne la même impression à l'observateur quelle que soit la position spatiale de la fenêtre à travers laquelle il observe cette texture. Par contre l'échelle d'observation doit être précisée. On peut le faire par exemple en précisant la taille de la fenêtre d'observation.

La notion de texture est liée à trois concepts principaux:

1- un certain ordre local qui se répète dans une région de taille assez grande.

2- cet ordre est défini par un arrangement structuré de ses constituants élémentaires.

3- ces constituants élémentaires représentent des entités uniformes qui se caractérisent par des dimensions semblables dans toute la région considérée.

Il existe un grand nombre de textures. On peut les séparer en deux classes: les textures structurées (macrotextures) et les textures aléatoires (microtextures).

Une texture qualifiée de structurée est constituée par la répétition d'une primitive à intervalle régulier. On peut différencier dans cette classe les textures parfaitement périodiques (carrelage, damier, etc.), les textures dont la primitives subit des déformations ou des changements d'orientation (mur de briques, grains de café, etc.).

Les textures qualifiées d'aléatoires se distinguent en général par un aspect plus fin (sable, herbe, etc.). Contrairement aux textures de type structurel, les textures aléatoires ne comportent ni primitive isolable, ni fréquence de répétition. On ne peut donc pas extraire de ces textures une primitive qui se répète dans l'image mais plutôt un vecteur de paramètres statistiques homogènes à chaque texture.

Dans tous les cas, ces objectifs nécessitent l'extraction d'un ou de plusieurs paramètres caractéristiques de cette texture. Nous désignerons ces paramètres sous le terme d'attributs texturaux (textural features) et l'ensemble qu'ils constituent sous le terme de descripteur de texture.

Certains de ces paramètres correspondent à une propriété visuelle de la texture (comme la directionnalité ou la rugosité). D'autres correspondent à des propriétés purement mathématiques auxquelles il est difficile d'associer une qualification perceptive.

Un recensement ainsi qu'une classification des termes de description des textures employés par les principaux auteurs pourront être trouvés dans [Rao et Loh] et [Rao et Loh, 1993].

Les attributs texturaux peuvent être obtenus à partir d'un ensemble assez vaste de différentes théories mathématiques. Citons notamment :

- Les attributs texturaux peuvent être obtenus à partir d'un ensemble assez vaste de différentes théories mathématiques. Citons notamment :
- Les attributs fondés sur des calculs statistiques effectués sur les niveaux de gris des pixels de l'image. C'est le cas des statistiques classiques, et des matrices de cooccurrences ou de longueurs de plages ainsi que les méthodes utilisant directement la fonction de covariance ou les statistiques d'ordre supérieur.
- Les attributs obtenus à la suite de transformations orthogonales appliquées aux images (transformées de Fourier, Ondelettes, etc.). Les attributs texturaux seront alors calculés dans des domaines différents de celui de la grille spatiale des luminances (domaine spectral par exemple).

- D'autres méthodes, basées par exemple sur la morphologie mathématique [F.pret, 1988] [Grat], les intégrales curvilignes [Barb, 1984], l'application de filtres [Fogel, 1989] [A.C.Bov] [A.K.JAI] [Kova et Vett].

Figure II.4 : Des textures différentes**[Sito 11]**

2.2.1. Les matrices de co-occurrences :

En 1973, Haralick [R.M.Ha] a proposé une méthode en se basant sur les matrices de co occurrences de niveaux de gris. La texture d'une image peut être interprétée comme la régularité d'apparition de couples de niveaux de gris selon une distance donnée dans l'image. La matrice de co-occurrences contient les fréquences spatiales relatives d'apparition des niveaux de gris selon quatre directions.
$(\theta = 0, \theta = \pi/4, \theta = \pi/2, \theta = 3\pi/4)$.

Une matrice de co-occurrences est définie au moyen d'une relation géométrique π entre deux pixels

(x_1, y_1) et (x_2, y_2).

La matrice de cooccurrences $P_{d,\theta}(i,j)$,q est carrée et de dimension $\Delta * \Delta$, Δ où est le nombre de niveaux de gris présents dans I. Les indices de la matrice de co occurrences sont donc les niveaux de gris de la texture étudiée.

On définit la matrice de co-occurrences $P_{d,\theta}$ par $P_{d,\theta} = (P_{d,\theta}(i,j))$

$P_{d,\theta}(i,j)$ représente le nombre de fois où un couple de points séparés par la distance d dans la direction a présenté les niveaux de gris et I_i I_j Pour obtenir θ de véritables fréquences relatives, il faut normaliser les éléments de la matrice en les divisant par le nombre total de paires de points élémentaires séparés par la distance d dans la direction dans toute l'image.

2.2.2. Transformée en ondelettes :

La transformée en ondelettes est à la base de nombreuses analyses de texture, telles que les filtres de Haar [Vail]. La description de texture à base d'ondelettes est utilisée pour la recherche d'images.

Pour avoir plus d'information sur les fondements mathématiques de la transformée en ondelettes, le lecteur peut se référer au livre [Vail]. Comme pour la transformée de Fourier, une présentation plus pédagogique et plus historique des ondelettes peut également être trouvée dans [Bur, 1995].L'approche continue des ondelettes pour un signal 2D est trop complexe pour être applicable rapidement sur des images. Pour résoudre ce problème, Mallat [Kova] considère l'analyse en ondelettes comme une décomposition du signal par une cascade de filtres, en utilisant une paire de filtres pour chaque niveau de résolution (un filtre passe-haut et un filtre passe-bas). Il propose ainsi la DWT (Discrete Wavelet Transform) qui permet d'obtenir une transformée rapide. Le choix de l'ondelette mère est alors remplacé par le choix du filtre. Pour calculer une transformée en ondelettes, on n'a alors besoin que des deux filtres : au lieu de calculer le produit scalaire de l'ondelette avec le signal, on réalise un produit de convolution du signal avec ces filtres**.**

Une des transformées en ondelettes les plus couramment employées en analyse d'images est la transformée de Haar, mais d'autres ondelettes sont aussi largement exploitées [Vail]. Les filtres de Haar sont fréquemment employés en apprentissage pour obtenir la description d'un objet (comme un visage ou une personne).

Conclusion :

Les attributs texturaux sont des attributs très importants pour la description de l'image et la reconnaissance des objets, cependant elles ne suffisent pas pour une bonne représentation du contenu de l'image, un autre attribut essentiel est la forme. Dans la suite nous allons introduire cet attribut et les différentes approches utilisées pour l'extraire.

2.3. Descripteurs de Formes :

La forme est un descripteur très important dans l'indexation des images. La forme désigne l'aspect général d'un objet, son contour. Nous présentons dans ce qui suit la méthode utilisé permettant de reconnaître une forme donnée dans une image. Itérée pour toutes les formes d'une image, cette méthode permet finalement de relever toutes les formes communes à deux images.

2.3.1. Les attributs géométriques de région :

Les attributs géométriques de forme permettent de distinguer les différents types de forme que peuvent prendre les objets d'une scène. Ils nécessitent une segmentation en région préalable de l'image. Ils sont ensuite calculés sur les différentes régions de l'image.

La surface relative (ou normalisée) d'une région $\Re_k$ de l'image I est le nombre de pixels contenus dans cette région par rapport au nombre total de pixels de l'image :

$$S_k = \frac{card(R_k)}{hauteur(I) * largeur(I)} \qquad (2.3.1)$$

Le centre de masse des pixels de la région est définie par:

$$P = (P_i, P_j) = \left(\frac{\sum_{i \in R_k} i / card(R_k)}{largeur(I)}, \frac{\sum_{j \in R_k} j / card(R_k)}{hauteur(I)} \right) \qquad (2.3.2)$$

La longueur du contour de la région est le nombre de pixels en bordure de la région:

$$l_k = card(contour(R_k)) \qquad (2.3.3)$$

La compacité traduit le regroupement des pixels de la région en zones homogènes et non trouées:

$$C_k = \frac{l_k^2}{S_k} \qquad (2.3.4)$$

Ces attributs très simples permettent d'obtenir des informations sur la géométrie des régions de l'image. Il existe d'autres attributs de forme, basés sur des statistiques sur les pixels des régions de l'image.

2.3.2. Les moments géométriques :

Les moments géométriques permettent de décrire une forme à l'aide de propriétés statistiques. Ils sont simples à manipuler mais leur temps de calcul est très long. Formule générale des moments:

$$m_{p,q} = \sum_{p=0}^{m} \sum_{q=0}^{n} x^p y^q f(x, y) \qquad (2.3.5)$$

L'ordre du moment est p + q. Le moment d'ordre 0 $m_{0,0}$ représente l'aire de la forme de l'objet. Les deux moments d'ordre 1 $m_{0,1}$ et $m_{1,0}$ associés au moment d'ordre 0 permettent de calculer le centre de gravité de l'objet. Les coordonnées de ce centre sont :

$$x_c = \frac{m_{1,0}}{m_{0,0}} \qquad y_c = \frac{m_{0,1}}{m_{0,0}} \qquad (2.3.6)$$

Il est possible de calculer à partir de ces moments l'ellipse équivalente à l'objet. Afin de calculer les axes de l'ellipse, il faut ramener les moments d'ordre 2 au centre de gravité :

$$m^g_{2,0} = m_{2,0} - m_{0,0}x_c^2 \quad m^g_{1,1} = m_{1,1} - m_{0,0}x_c y_c \quad m^g_{0,2} = m_{0,2} - m_{0,0}y_c^2$$

Puis on détermine l'angle α d'inclinaison de l'ellipse

$$\alpha = \frac{1}{2}\arctan\frac{2m^g_{1,1}}{m^g_{2,0} - m^g_{0,2}} \qquad (2.3.7)$$

À partir des moments géométriques, Hu [Hu, 1962] a introduit sept invariants aux translations, rotations et changement d'échelle, appelés moments de Hu.

$$\begin{aligned}
M_1 &= \mu_{20} + \mu_{02}. \qquad (2.3.8)\\
M_2 &= (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2.\\
M_3 &= (\mu_{30} - 3\mu_{21})^2 + (3\mu_{21} - \mu_{03})^2.\\
M_4 &= (\mu_{30} + \mu_{21})^2 + (\mu_{21} + \mu_{03})^2.\\
M_5 &= (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})\left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2\right] +\\
&\quad (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})\left[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2\right].\\
M_6 &= (\mu_{20} - \mu_{02})\left[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2\right] + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{03} + \mu_{21}).\\
M_7 &= (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})\left[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})\right] -\\
&\quad (\mu_{30} - 3\mu_{21})(\mu_{12} + \mu_{03})\left[3(\mu_{30} + \mu_{12})^2 - (\mu_{12} + \mu_{03})^2\right].
\end{aligned}$$

Les moments de Hu offrent d'excellents attributs invariants en translation, rotation et changement d'échelle pour décrire une image. Cependant leur calcul est relativement long et ils sont très sensibles au bruit, ce qui peut s'avérer être un gros inconvénient dans un système de recherche d'images.

2.3.3. Transformée de Hough :

Soit $\Re^n$ l'espace image, et ξ un ensemble de N points sélectionnés par un prétraitement :

$$\xi = \{M_i, i = 1...\overline{N}\} \in \Re^n \qquad (2.3.9)$$

Un point M de $\Re^n$ est repéré par ses coordonnées x.

Soit p$\Omega \subset \Re^p$ un espace de paramètres et F une famille de courbes dans n $\Re^n$ paramétrée par a :

$$F = \left\{ \left\{ x : f(x,a) = 0,\ x \in \Re^n \right\},\ a \in \Omega \right\}. \qquad (2.3.10)$$

On appelle transformation de Hough associée à la famille F une transformation qui fait correspondre à l'ensemble x.

Conclusion :

Les formes représentent un descripteur puissant pour décrire les objets contenu dans l'image.

3. Mesures de similarité :

Afin de déterminer les descripteurs d'images, on peut calculer la valeur de similarité entre les images de base et notre image requête. Au lieu d'un appariement exact, la recherche d'images par le contenu calcule des similarités visuelles entre une image requête et les images de la base d'images. En conséquent, le résultat d'une recherche n'est pas une seule image mais une liste d'images ordonnées selon leur degré de similitude avec l'image requête. Plusieurs mesures de similarités ont été proposées dans la littérature. Les différentes mesures de similarité influencent les performances de recherche des systèmes de recherche par le contenu.

Plusieurs mesures de similarité sont basées sur la distance Lp entre deux points. Pour deux points donnés (x, y) dans $\Re^k$ la distance Lp est définie par :

$$L_p = \left(\sum_{i=1}^{k} |x_i - y_i|^p\right)^{1/p} \qquad (2.3.11)$$

Ou P=1,2 ou ∞.

Le choix de la mesure de similarité la plus appropriée dépend du niveau d'abstraction de la représentation de l'image : images brutes (pixels) ou attributs visuels.

- **Images brutes (pixels) :**

Au plus bas niveau d'abstraction, les images sont tout simplement des agrégations de pixels. La comparaison entre les images, est réalisée pixel par pixel, et les mesures de similarité couramment utilisées comprennent : le coefficient de corrélation, la somme des valeurs absolue des différences

(SVAD), la distance des moindres carrés, et l'information mutuelle. La comparaison au niveau des pixels est très spécifique et, par conséquent, n'est utilisée que lorsque des appariements relativement précis sont nécessaires.

- **Attributs visuels :**

Les attributs visuels sont des valeurs numériques de données extraites des images ou des objets dans les images, tels que la couleur, la forme et la texture. Plusieurs mesures de similarité sont couramment utilisées pour la comparaison d'attributs: la distance euclidienne, la distance de Minkowsky et la distance d'intersection.

3.1. Les méthodes de calcul :

Ci-après les distances les plus utilisées pour comparer des images considérées comme vecteurs ou comme distributions statistiques. [Houa, 2010]

3.1.1. Distance de Mahalanobis :

Cette distance prend en compte la corrélation entre les distributions des classes. Où **C** est la matrice de covariance. Dans les cas où les dimensions des caractéristiques sont indépendantes, C ne comporte que dès les variances et la distance de Mahalanobis se simplifie sous la forme :

$$D_M = \sqrt{(f_1 - f_2)^T C^{-1} (f_1 - f_2)}. \qquad (3.1.1)$$

Si **C** est la matrice identité, est la distance euclidienne.

$$D_M = \frac{\sum (f_1(i) - f_2(i))^2}{c_i} \qquad (3.1.2)$$

3.1.2. Intersection d'histogrammes :

Cette mesure est l'une des premières distances utilisée dans la recherche d'image par le contenu. Elle a été proposée par Swain et Ballard mesurant la partie commune entre deux histogrammes. Etant donné deux histogrammes h_1 et h_2

$$D_{Intersec} = \frac{\sum_i min(h_1(i), h_2(i))}{\sum_i h_2(i)} \qquad (3.1.3)$$

Deux images présentant une intersection normalisée d'histogrammes proche de 1 sont considérées comme similaires. Cette mesure n'est pas une métrique parce que non symétrique. Cependant il en existe des versions symétriques telles que celle proposée par Smith [Y.Rub].

3.1.3. Earth Mover Distance (EMD) :

EMD consiste à minimiser le coût de transformation d'une distribution en une autre sous certaines contraintes de déplacement des classes de descripteurs. EMD requiert une optimisation linéaire.

$$D_{EMD} = \frac{\sum_{ij} g_{ij} d_{ij}}{\sum_{ij} g_{ij}} \qquad (3.1.4)$$

Où d_{ij} représente la disimilarité entre deux indices (i, j) et g_{ij} est le flot optimal entre deux distributions dont le coût total est:
Le coût est minimisé sous les contraintes suivantes :

$$D_{EMD} = \sum_{ij} g_{ij} d_{ij}$$

$$g_{ij} \geq 0, \forall i, j$$

$$\sum_{i} g_{ij} \leq f_2(j), \forall j$$

$$\sum_{j} g_{ij} \leq f_1(i), \forall i$$

$$\sum_{i} \sum_{j} g_{ij} = \min\left(f_1(i), f_2(j)\right)$$

EMD prétend également mimer la vision humaine

3.1.4. Distance de Minkowski :

La distance de Minkowski est une famille de distances vectorielles. Soit deux vecteurs de caractéristiques, elle s'exprime par :

$$d^p(f_1, f_2) = \left(\sum_{i=1}^{n} |f_1(i) - f_2(i)|^p\right)^{1/p} \qquad (3.1.5)$$

P est le facteur de Minkowski et n la dimension de l'espace caractéristique. La distance Euclidienne est un cas particulier de cette distance où p=2, de même que la distance de Manhatan (p=1).

3.1.5. Distance quadratique :

La distance de Minkowski traite les éléments du vecteur de caractéristique d'une manière équitable. La distance quadratique en revanche favorise les éléments les plus ressemblants. Sa forme générale est donné par : $D_Q = \sqrt{(f_1 - f_2)^T A (f_1 - f_2)}$ où $A = [a_{ij}]$ est la matrice de similarité. Représente la distance entre deux éléments des vecteurs $f_1 et\ f_2$. Hafner et al [al] propose la formule suivante pour construire la matrice A.

$$a_{ij} = 1 - \frac{d_{ij}}{\max(d_{ij})} \qquad (3.1.6)$$

Les propriétés de cette distance la rendraient proche de la perception humaine de la couleur, ce qui en fait une métrique attractive pour les systèmes de Recherche d'images couleur par le contenu.

3.1.6. Distance de Bhattacharya :

La distance de Bhattacharya exploite la séparabilité entre deux distributions gaussiennes représentées par leur covariance : Σ

$$D_B = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) + \frac{1}{2} ln \frac{\det(\Sigma)}{\sqrt{\det(\Sigma_1)\det(\Sigma_2)}} \qquad (3.1.7)$$

Où $\Sigma = 0.5 \times (\Sigma_1 + \Sigma_2)$ La séparabilité entre classes est estimée par la distance des moyennes et des matrices de covariance de chaque classe.

3.1.7. Distance de Kullback Leiber (KL) :

La divergence de Kullback Leiber exprime l'entropie relative de deux distributions :

$$D_{KL} = \sum_i f_1(i) log \frac{f_1(i)}{f_2(i)} \qquad (3.1.8)$$

3.1.8. Divergence de Jeffrey (JD) :

La divergence de Jeffrey est défini par :

$$D_{JD} = \sum_i f_1(i) log \frac{f_1(i)}{\hat{f}_i} + f_2(i) log \frac{f_2(i)}{\hat{f}_i} \qquad (3.1.9)$$

(3.1.10)

Où $\hat{f}_i = (f_1(i) + f_2(i))/2$ A la différence de la mesure KL. JD est symétrique et plus stable.

3.1.9. Distance de Kolmogorov Smirnov :

Cette distance est appliquée aux distributions cumulées $f^c(i)$

$$D_{KS} = \max_i |f_1^c(i) - f_2^c(i)| \qquad (3.1.11)$$

3.1.10. Distance de Cramer Von Mises :

La distance de Cramer Von Mises s'applique également sur des distributions cumulées, elle est définie par :

$$D_{CVM} = \sum_i (f_1^c(i) - f_2^c(i))^2 \qquad (3.1.12)$$

4. Conclusion :

Le choix des descripteurs pour un système de recherche d'images par contenu est important, dans le sens où, ce choix influe sur les résultats attendus. Cependant, d'une part il n'y a pas d'attributs universels, et d'autre part le choix des descripteurs dépend fortement de la base d'image à utiliser et des connaissances à priori qu'on peut avoir sur la base.

Chapitre 3: Apprentissage profond et vision par ordinateur

1. Introduction:

Le Deep Learning ou apprentissage profond est un type d'intelligence artificielle, dérivé du machine Learning qui a été développé dans le but de créer des algorithmes capables d'apprendre et de s'améliorer de manière autonome, contrairement à la programmation où la machine se contente d'exécuter à la lettre des règles prédéterminées. L'apprentissage en profondeur utilise une succession de couches d'unités de traitement non linéaire pour pouvoir extraire ou transformer les caractéristiques des données. La sortie d'une couche sert d'entrée de la couche suivante. Les algorithmes de l'apprentissage profond peuvent être supervisés et servir à classer les données, ou non supervisés et aider à effectuer une analyse de modèle. L'algorithme de Deep Learning absorbe des quantités de données énormes par rapport aux autres algorithmes d'apprentissages machine utilisés et développés actuellement, et il a été capable de battre les humains dans certaines tâches cognitives. Par exemple, la reconnaissance faciale par ordinateur et la reconnaissance vocale ont connu des progrès significatifs et cela grâce aux approches d'apprentissage approfondies. Parmi ces approches d'apprentissage approfondies, on compte les réseaux de neurones artificiels sur lesquels reposent le Deep Learning et ainsi certaines technologies comme la reconnaissance d'image ou la vision robotique. Les réseaux de neurones artificiels sont inspirés par les neurones du cerveau humain. Ils sont constitués à base de plusieurs neurones artificiels connectés entre eux. Plus le nombre de neurones est considérable, plus le réseau est profond.

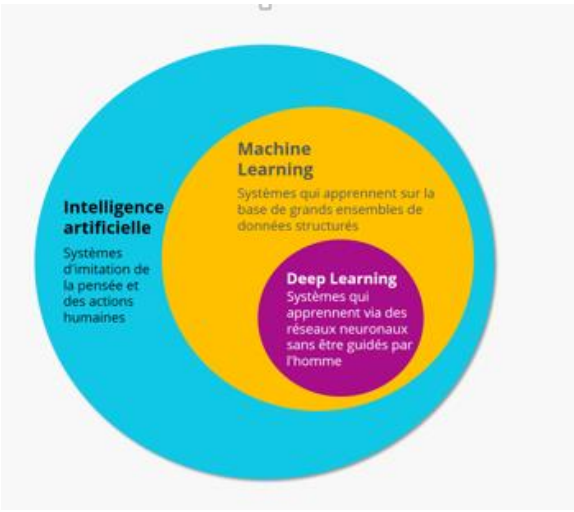

Figure I. 1: La relation entre l'IA, Machine Learning et Deep Learning [Sito 12]

2. Les réseaux de neurones artificiels :

2.1 Définition :

Un réseau de neurones artificiels, ou (Artificiel Neural Network en anglais), est un système informatique matériel et / ou logiciel s'inspirant du fonctionnement du cerveau humain pour apprendre. Il s'agit d'une variété de technologie Deep Learning, qui fait elle-même partie de la sous-

catégorie d'intelligence artificielle et du Machine Learning. Ce genre de réseau est défini par un ensemble de couches de neurones qui sont fortement interconnectées entre elles.

2.2 .Architecture d'un réseau de neurones artificiel :

En général un réseau de neurones est constitué d'un ensemble de couches successives dont chacune prend ses entrées sur les sorties de la précédente c.-à-d. que cet ensemble est entièrement connectée. Chaque couche est un ensemble de neurones n'ayant pas de connexion entre eux et qui reçoivent des informations numériques en provenance de neurones voisins. L'ensemble de couches est composé d'une couche d'entrée qui lit les valeurs d'entrées, une couche de sortie qui fournit les résultats du système et entre ces deux se cache une à plusieurs couches dites cachées qui participent au transfert. Comme le montre la figure 2 suivante

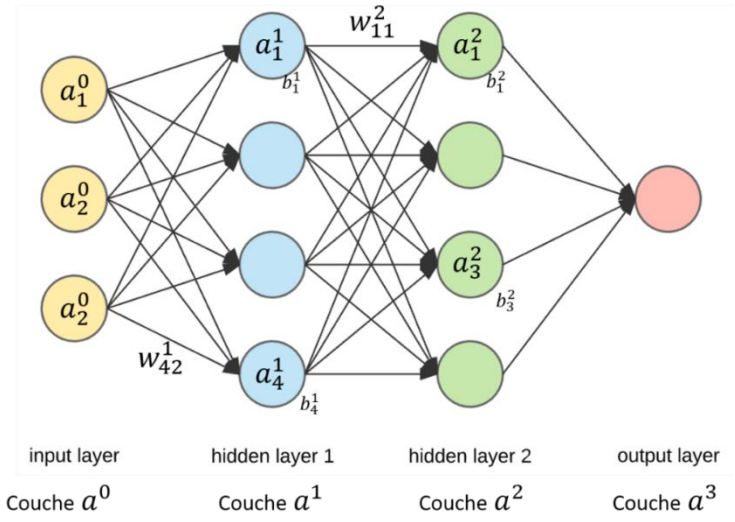

Figure III.2: Architecture de base d'un réseau de neurones artificiel [Sito 13]

2.3 Types des réseaux de neurones :

Il existe différents types de réseaux de neurones, et ils sont classés en fonction du nombre de nœuds cachées du modèle ou encore du nombre d'entrées et de sorties de chaque nœud. La propagation des informations entre les différents neurones peut varier et dépend du type de réseaux de neurones.

- **Les réseaux de neurones dit "feed-forward" (à propagation avant):** c'est la variante la plus simple, l'information ne se déplace que dans une seule direction, elle traverse directement l'entrée aux nœuds de traitement (couches cachées) puis aux sorties, avec absence de cycle ou de boucle dans le réseau.

● **Les réseaux de neurones récurrents**: Ce mode d'apprentissage est un peu plus complexes, ils comportent au moins un cycle dans leurs structures, ils sauvegardent les résultats produits par les nœuds de traitement et nourrissent le modèle à l'aide de ces résultats. Parmi ses applications, on trouve: la reconnaissance automatique de formes, la traduction automatique de la parole...etc.

● **Les réseaux de neurones convolutifs** : leur fonctionnement est inspiré par un processus biologique qui est le cortex visuel des animaux, les données à traiter.

3. Réseau de neurones convolutifs (CNN) :

3.1 Définition :

Les réseaux de neurones convolutifs sont un type spécialisé de réseaux de neurones artificiels qui utilisent une opération mathématique appelée convolution à la place de la multiplication matricielle générale dans au moins une de leurs couches.[Ian] Ils sont spécifiquement conçus pour traiter les données de pixels et sont utilisés dans la reconnaissance et le traitement d'images.

3.2 Architecture de base de convolutional neural network :

Le réseau neuronal convolutif se compose d'une couche d'entrée, de couches cachées et d'une couche de sortie. Dans tout réseau de neurones à anticipation, toutes les couches intermédiaires sont dites cachées car leurs entrées et sorties sont masquées par la fonction d'activation et la convolution finale. Dans un réseau neuronal convolutif, les couches cachées comprennent des couches qui effectuent des convolutions. Cela inclut généralement une couche qui effectue un produit scalaire du noyau de convolution avec la matrice d'entrée de la couche. Ce produit est généralement le produit interne de Frobenius et sa fonction d'activation est généralement ReLU. Lorsque le noyau de convolution glisse le long de la matrice d'entrée pour la couche, l'opération de convolution génère une carte de caractéristiques, qui à son tour contribue à l'entrée de la couche suivante. Viennent ensuite d'autres couches telles que les couches de regroupement, les couches entièrement connectées et les couches de normalisation.

- **Couche convolutive :**

Dans un CNN, l'entrée est un tenseur de forme : (nombre d'entrées) x (hauteurs d'entrée) x (largeurs d'entrée) x (canaux d'entrée). Après avoir traversé une couche convolutive, l'image devient abstraite en une carte de caractéristiques, également appelée carte d'activation, avec la forme : (nombre d'entrées) x (hauteurs de la carte de caractéristiques) x (largeurs de la carte de caractéristiques) x (canaux de la carte de caractéristiques).

Les couches convolutionnelles convoluent l'entrée et transmettent son résultat à la couche suivante. Ceci est similaire à la réponse d'un neurone du cortex visuel à un stimulus spécifique. [LeNet] Chaque neurone convolutif traite les données uniquement pour son champ récepteur. Bien que des réseaux de neurones à anticipation entièrement connectés puissent être utilisés pour apprendre des fonctionnalités et classer des données, cette architecture est généralement peu pratique pour des entrées plus importantes telles que des images hautes résolution. Cela nécessiterait un nombre très élevé de neurones, même dans une architecture peu profonde, en raison de la grande taille d'entrée des images, où chaque pixel est une caractéristique d'entrée pertinente. Par exemple, une couche entièrement connectée pour une (petite) image de taille 100 x 100 à 10 000 poids pour chaque neurone de la deuxième couche. Au lieu de cela, la convolution réduit le nombre de paramètres libres, permettant au réseau d'être plus profond. [Habi] Par exemple, quelle que soit la taille de l'image, l'utilisation d'une région de mosaïque 5 x 5, chacune avec les mêmes poids partagés, ne nécessite que 25 paramètres a prenables. L'utilisation de poids régularisés sur moins de paramètres évite les problèmes de gradients de fuite et d'explosion de gradients observés lors de la rétro propagation dans les réseaux de neurones traditionnels. De plus, les réseaux de neurones convolutifs sont idéaux pour les données avec une topologie en forme de grille (telles que les images) car les relations spatiales entre les caractéristiques distinctes sont prises en compte lors de la convolution et/ou de la mise en commun.

- **Mise en commun des couches :**

Les réseaux convolutifs peuvent inclure des couches de regroupement locales et/ou globales ainsi que des couches convolutives traditionnelles. Les couches de regroupement réduisent les dimensions des données en combinant les sorties des grappes de neurones d'une couche en un seul neurone de la couche suivante. La mise en commun locale combine de petits clusters, des tailles de mosaïque telles que 2 x 2 sont couramment utilisées. La mise en commun globale agit sur tous les neurones de la carte des caractéristiques. [Cire,2013][Kriz] Il existe deux types courants de mise en commun couramment utilisés : le maximum et la moyenne. La mise en commun maximale utilise la valeur maximale de chaque groupe local de neurones dans la carte des caractéristiques, [Yama] [Cire, 2012] tandis que la mise en commun moyenne prend la valeur moyenne.

- **Couches entièrement connectées :**

Des couches entièrement connectées connectent chaque neurone d'une couche à chaque neurone d'une autre couche. C'est la même chose qu'un réseau de neurones perceptron

multicouche traditionnel (MLP). La matrice aplatie passe par une couche entièrement connectée pour classer les images

- **Champ réceptif :**

Dans les réseaux de neurones, chaque neurone reçoit une entrée d'un certain nombre d'emplacements dans la couche précédente. Dans une couche convolutive, chaque neurone ne reçoit d'entrée que d'une zone restreinte de la couche précédente appelée champ récepteur du neurone. Typiquement, la zone est un carré (par exemple 5 par 5 neurones). Alors que, dans une couche entièrement connectée, le champ récepteur est toute la couche précédente. Ainsi, dans chaque couche convolutive, chaque neurone reçoit une entrée d'une plus grande zone d'entrée que les couches précédentes. Cela est dû à l'application répétée de la convolution, qui prend en compte la valeur d'un pixel, ainsi que ses pixels environnants. Lors de l'utilisation de calques dilatés, le nombre de pixels dans le champ récepteur reste constant, mais le champ est moins peuplé au fur et à mesure que ses dimensions augmentent lors de la combinaison de l'effet de plusieurs calques.

- **Poids :**

Chaque neurone d'un réseau de neurones calcule une valeur de sortie en appliquant une fonction spécifique aux valeurs d'entrée reçues du champ récepteur dans la couche précédente. La fonction appliquée aux valeurs d'entrée est déterminée par un vecteur de poids et un biais (généralement des nombres réels). L'apprentissage consiste à ajuster itérativement ces biais et ces poids.

Le vecteur de poids et le biais sont appelés filtres et représentent des caractéristiques particulières de l'entrée (par exemple, une forme particulière). Une caractéristique distinctive des CNN est que de nombreux neurones peuvent partager le même filtre. Cela réduit l'empreinte mémoire car un seul biais et un seul vecteur de poids sont utilisés dans tous les champs récepteurs qui partagent ce filtre, par opposition à chaque champ récepteur ayant son propre biais et sa propre pondération vectorielle.

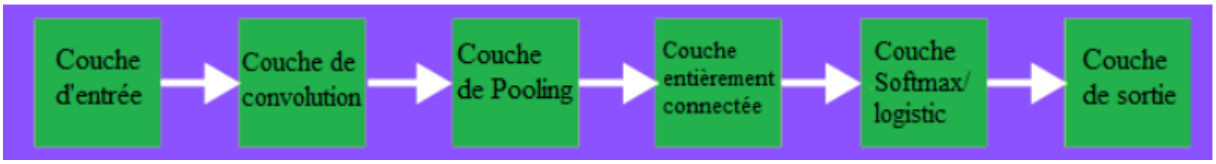

Figure III. 3 : **Architecture de base de convolutional neural network [Sito 14]**

4. Architectures d'apprentissage en profondeur pour la vision par ordinateur:

Les performances et l'efficacité d'un CNN sont déterminées par son architecture. Cela inclut la structure des couches, la façon dont les éléments sont conçus et les éléments présents dans chaque couche. De nombreux CNN ont été créés, mais voici quelques-unes des conceptions les plus efficaces.

➢ Alex Net (2012) :

AlexNet est une architecture basée sur l'ancienne architecture LeNet. Il comprend cinq couches convolutionnelles et trois couches entièrement connectées. AlexNet utilise une structure à double pipeline pour permettre l'utilisation de deux GPU pendant la formation.La principale différence entre AlexNet et les architectures précédentes est son utilisation d'unités linéaires rectifiées (ReLU) au lieu des fonctions d'activation sigmoïde ou Tanh qui étaient utilisées dans les réseaux de neurones traditionnels. ReLU est plus simple et plus rapide à calculer, permettant à AlexNet de former des modèles plus rapidement. **[Run.ai]**

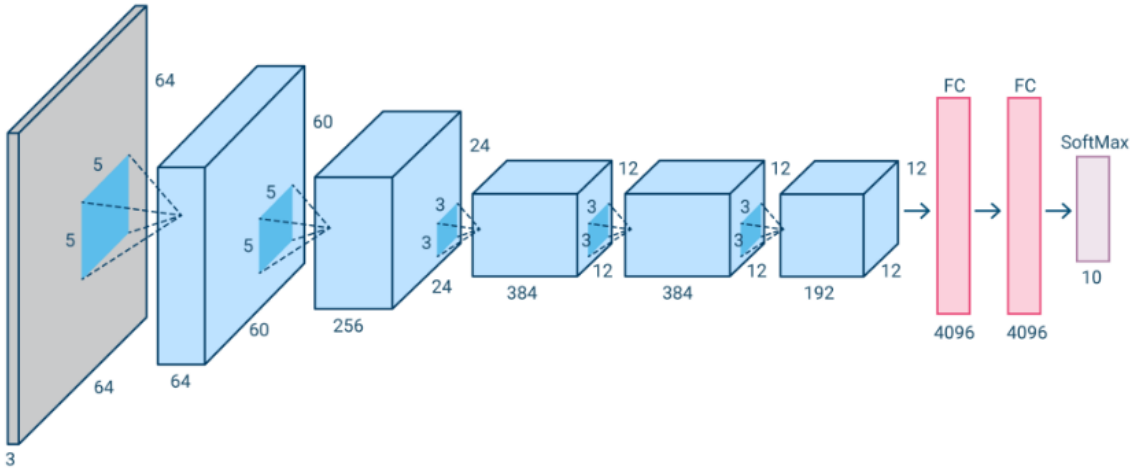

Figure III. 4 :Alex Net (2012) [Sito 14]

➢ Google Net (2014) :

GoogleNet, également connu sous le nom d'Inception V1, est basé sur l'architecture LeNet. Il est constitué de 22 couches constituées de petits groupes de convolutions, appelés « modules d'inception ». Ces modules de démarrage utilisent la normalisation par lots et RMSprop pour réduire le nombre de paramètres que GoogleNet doit traiter. RMSprop est un algorithme qui utilise des méthodes de taux d'apprentissage adaptatif. [Run.ai]

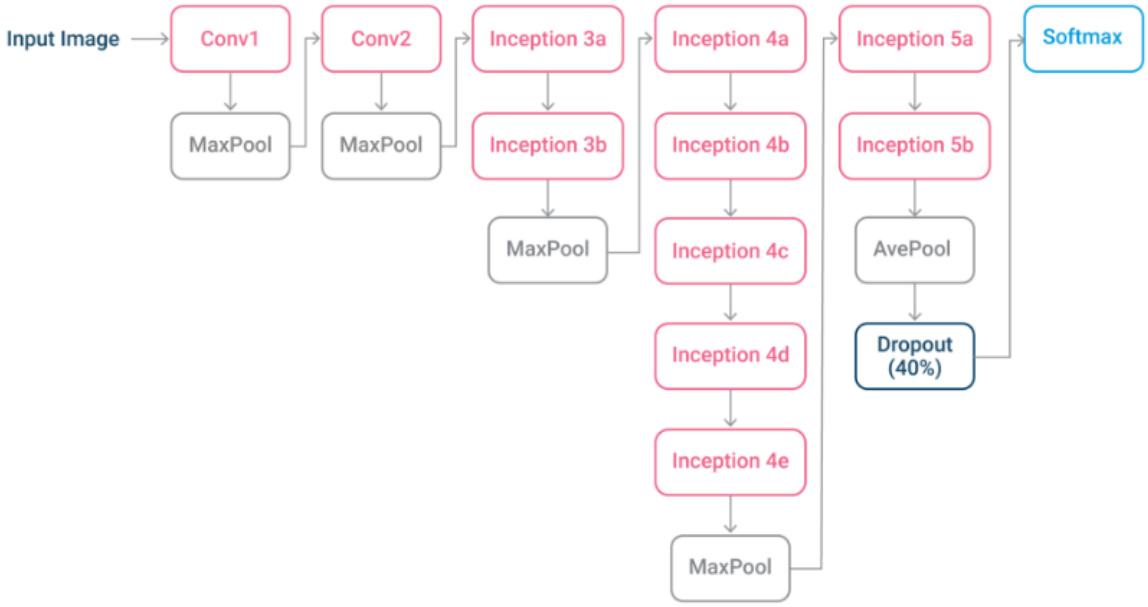

Figure III. 5: Googlenet [Siot 14]

➢ VGGnet (2014) :

VGG 16 est une architecture à 16 couches (certaines variantes avaient 19 couches). VGGNet a des couches convolutives, une couche de regroupement, quelques couches convolutionnelles supplémentaires, une couche de regroupement, plusieurs autres couches de conversion, etc.

VGG est basé sur la notion d'un réseau beaucoup plus profond avec des filtres plus petits - il utilise des convolutions 3 × 3 tout au long, ce qui est la plus petite taille de filtre convolution qui ne regarde que certains des pixels voisins. Il utilise de petits filtres en raison de moins de paramètres, ce qui permet d'ajouter plus de couches. Il a le même champ récepteur effectif que si vous aviez une couche convolutive 7 × 7. [Run.ai]

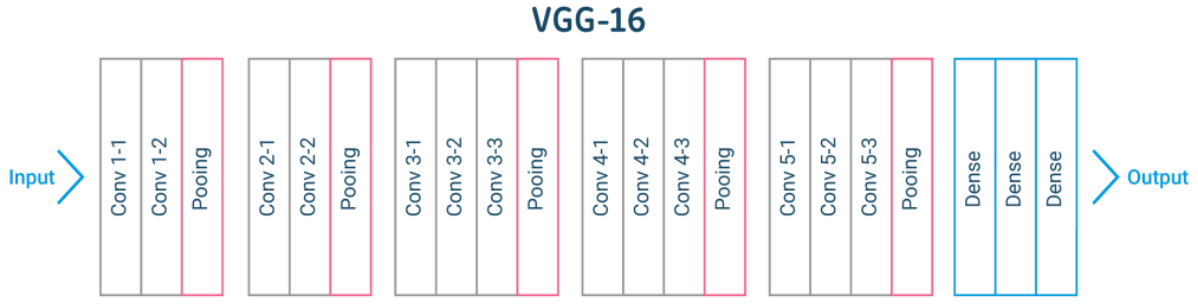

Figure III. 6 : l'architecture de VGG-16 [Sito 14]

➢ ResNet (2015) :

ResNet, abréviation de Residual Neural Network, est une architecture conçue pour avoir un grand nombre de couches - les architectures généralement utilisées vont de ResNet-18 (avec 18 couches) à ResNet-1202 (avec 1202 couches). Unités ou "connexions de saut" qui lui permettent de transmettre des informations à des couches convolutives ultérieures. ResNet utilise également la normalisation par lots pour améliorer la stabilité du réseau. [Run.ai]

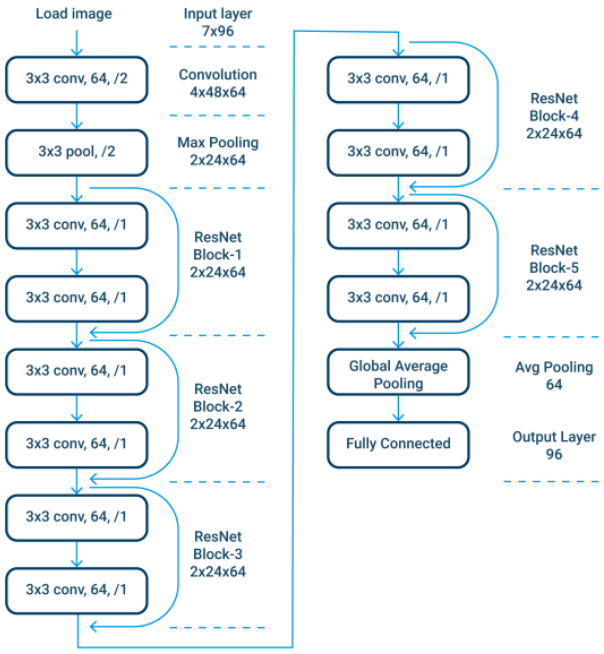

Figure III. 7: l'architecture ResNet(2015) [Sito 15]

➢ Xception (2016) :

Xception est une architecture basée sur Inception, qui remplace les modules d'inception par des convolutions séparables en profondeur (convolution en profondeur suivie de convolutions ponctuelles). Il fonctionne en capturant d'abord les corrélations cartographiques inter-entités, puis les corrélations spatiales. Cela permet une utilisation plus efficace des paramètres du modèle. [Run.ai]

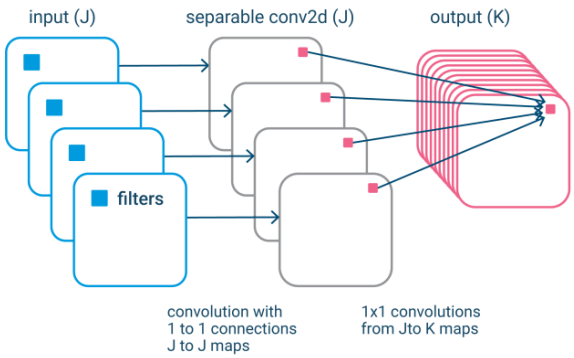

Figure III. 8 : l'architecture Xception (2016) [Sito 14]

➢ ResNeXt-50 (2017)

ResNeXt-50 est une architecture basée sur des modules avec 32 chemins parallèles. Il utilise la cardinalité pour réduire les erreurs de validation et représente une simplification des modules de démarrage utilisés dans d'autres architectures. [Run.ai]

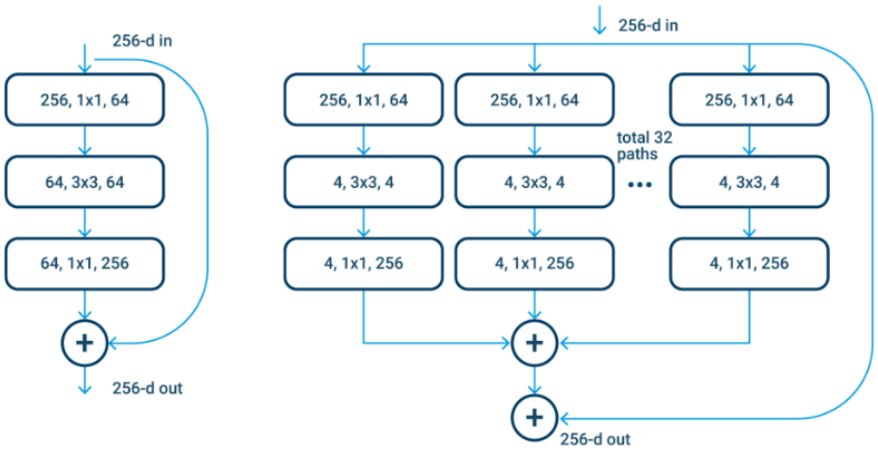

Figure III. 9 : l'architecture ResNeXt-50 [Sito 14]

5. Utilisations de l'apprentissage en profondeur dans la vision par ordinateur :

- **La reconnaissance faciale :**

Un algorithme de Deep Learning apprend à détecter sur une photo les caractéristiques du visage tels que: les yeux, la bouche, le nez...etc. Et cela en fournissant à l'algorithme un ensemble d'images du visage qui vont être utilisées pour l'entraînement, et à force de les entraîner, il sera capable de détecter un visage sur une image.

- **La détection d'objets :**

Les algorithmes de détection d'objets sont capables maintenant d'identifier au pixel près un élément ou une personne sur une image qui contient beaucoup d'éléments (images complexes)

- **Segmentation sémantique :**

La segmentation sémantique, également connue sous le nom de segmentation d'objet, est similaire à la détection d'objet, sauf qu'elle est basée sur les pixels spécifiques liés à un objet. Cela permet aux objets image d'être définis avec plus de précision et ne nécessite pas de cadres de délimitation. La segmentation sémantique est souvent effectuée à l'aide de réseaux entièrement convolutionnels (FCN) ou U-Nets.

Une utilisation populaire de la segmentation sémantique est la formation de véhicules autonomes. Avec cette méthode, les chercheurs peuvent utiliser des images de rues ou de passages avec des limites définies avec précision pour les objets.

- **Pose estimation :**

L'estimation de la pose est une méthode utilisée pour déterminer où se trouvent les articulations dans une image d'une personne ou d'un objet et ce que le placement de ces articulations indique. Il peut être utilisé avec des images 2D et 3D. L'architecture principale utilisée pour l'estimation de la pose est PoseNet, qui est basée sur les CNN.

L'estimation de pose est utilisée pour déterminer où des parties du corps peuvent apparaître dans une image et peut être utilisée pour générer des positions ou des mouvements réalistes de figures humaines. Souvent, cette fonctionnalité est utilisée pour la réalité augmentée, la mise en miroir de mouvements avec la robotique ou l'analyse de la marche.

6. Conclusion :

Les réseaux de neurones offrent un cadre de travail souple et robuste dont les intérêts ont été largement démontrés expérimentalement. Ils ont permis de franchir des étapes importantes pour le développement de l'apprentissage automatique, notamment le traitement de l'image où des performances de systèmes automatiques sont comparables à celles obtenues par les humains dans certaines conditions. Les très grands corpus rendus disponibles récemment par exemple le Yahoo news feeds dataset9, le Google audio set10 et le Google video set11 couplés aux puissantes capacités de modélisation des réseaux profonds qui évoluent rapidement, laissent présager l'exploration de nouveaux domaines de recherche ainsi que de nouvelles applications intéressante.

Chapitre 04: Contribution pratique

1. Introduction :

Dans ce chapitre, nous allons présenter la méthode utilisée, parmi toutes les méthodes existantes du deepLearning dans la recherche de l'image par le contenu (CBIR). Pour l'extraction des features de toutes les images nous avons utilisé deux descripteurs d'image l'architecture VGG-16 CNN . Dans ce chapitre, nous allons présenter la méthode utilisée, les résultats obtenus et analyse de ces résultats. D'abord nous expliquons le principe de la méthode utilisé VGG-16 et aussi les autres techniques utilisées pour la réalisation du CBIR. A l a fin nous présentons la méthodologie d'évaluation des résultats donnés comme la distance utilisé, la base d'images et performance des résultats.

2. Méthodes implantées :

2.1. VGG-16 :

Dans cette phase nous utilisons VGG-16 comme l'architecture, cette phase est divisée en deux parties : apprentissage et test les étapes sont les suivantes :

Etape1 : Apprentissage

- Extrait les caractéristiques de toutes les images d'entraînement :
  - Redimensionner l'image (224 x 224 pixels).
  - Convertir l'espace colorimétrique de l'image (en RGB).
  - Extrait les caractéristiques.

Etape2 : Test

- Cette partie lance le test du modèle.
  - Tester une image de requête pour récupérer des images similaires.
  - Insertion de l'image requête.
  - Extraire ses caractéristiques.
  - Calculer la similarité (Distance) entre images.
  - Extraire 30 images qui ont la distance la plus faible.

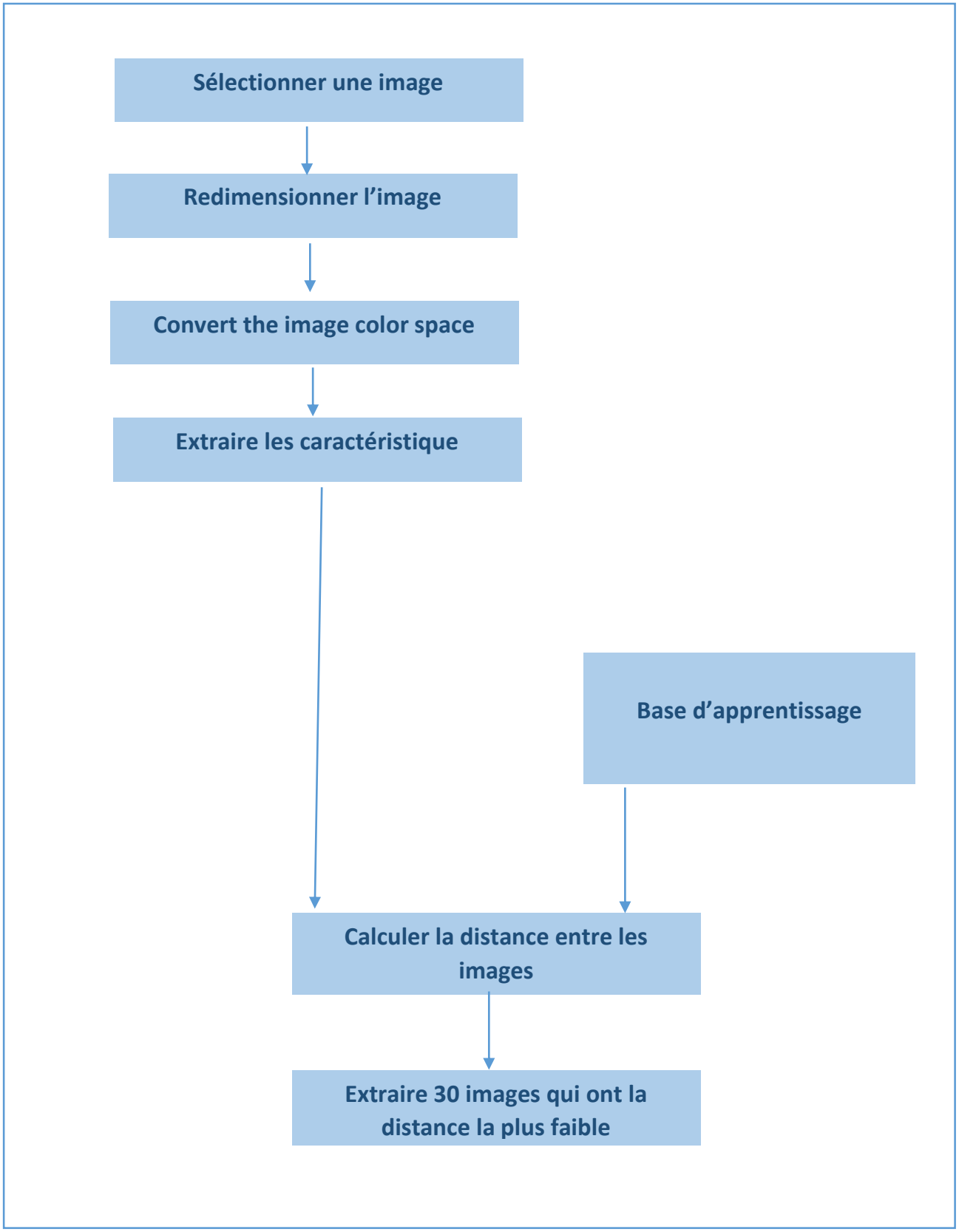

Figure IV. 1 : processus méthode implimentée

3. Base de données:

3.1. Présentation de la base COREL10-k :

Pour la base de données de notre travail nous avons utilisé la base d'images COREL 10-K, cette base de référence est constituée de 10000 images qui sont catégorisées manuellement en 80 classes, chacune contient plus de 100 images. On n'a utilisé que 10 classes pour le paramétrage des tests (les classes utilisé seront affiché dans la figure IV. 2). Les images sont en format JPG, en couleur de taille (256*384) pixel ou (384*256) pixel.

La base de données est disponible dans le site web suivant :

https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval

Figure IV. 2: Exemple d'images de la base Wang: les classes utilisées. [Sito 2]

4. Résultats et analyse:

4.1. Environnement de développement:

Pour l'implémentation du système que nous avons utilisé Python 3.10.4.

4.2. Mesure d'évaluation et performance:

Les mesures de performances, nous avons utilisé la précision qui est les critères le plus utilisé :

$$Précision = \frac{le\ nombre\ d'images\ bien\ classé}{le\ nombre\ totale\ d'images\ testé} \qquad (4.2.1)$$

4.3. Mesure de distance :

Nous avons utilisé la distance: Euclidienne.

4.4. Résultats:

➢ Exemple des requêtes :

Figure IV.3. Exemple d'une query (Rose)

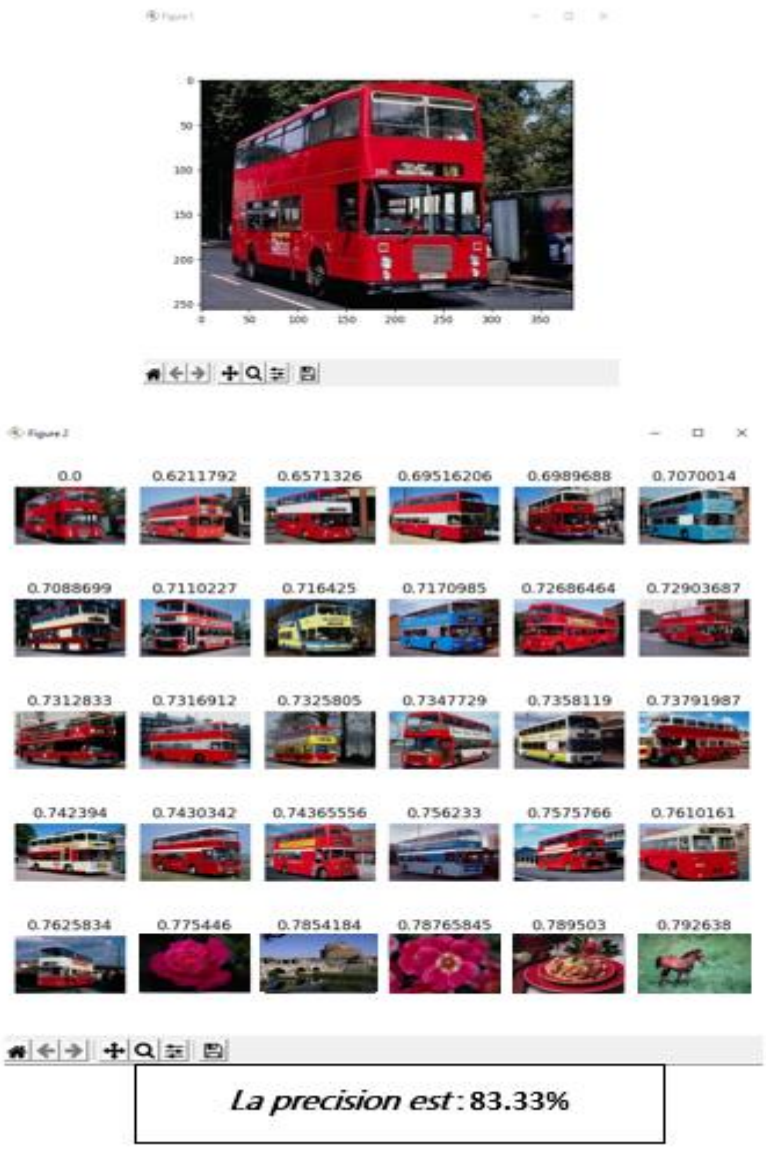

Figure IV.4. Exemple d'un query (Bus)

Figure IV.5. Exemple d'un query (Afrique)

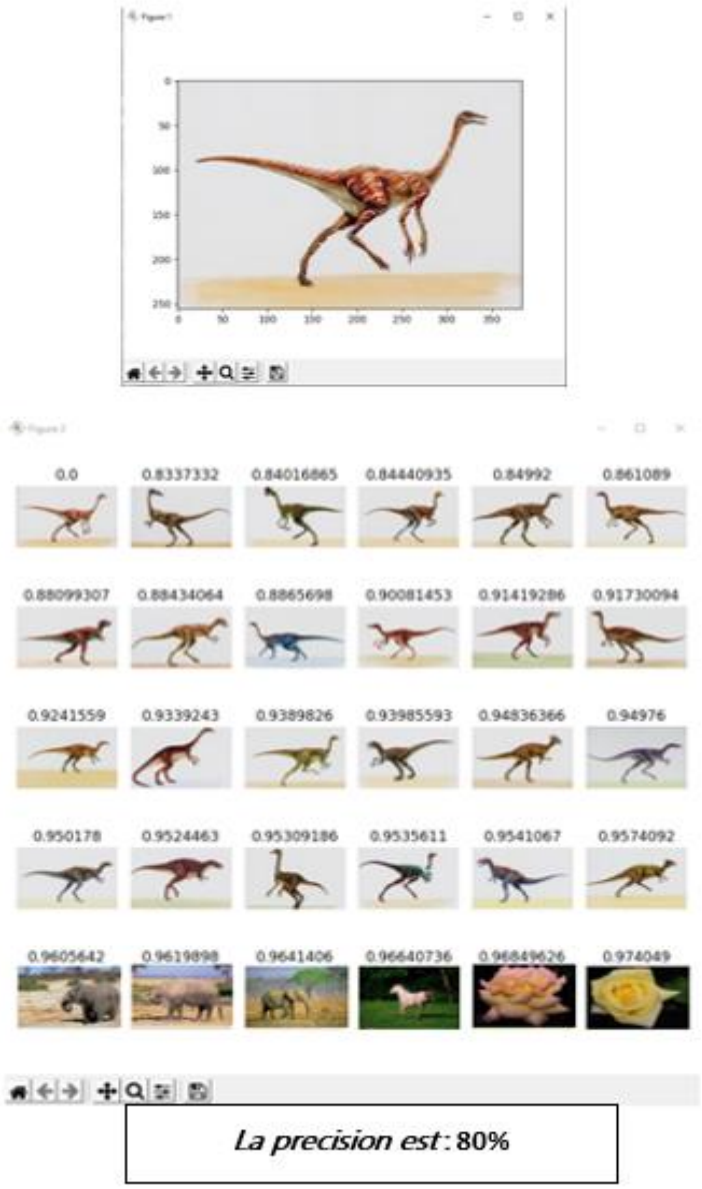

Figure IV.6. Exemple d'une query (Dinosaur)

Le Pourcentage de précision obtenu dans cette méthode pour chacune des 10 classes choisies est présenté dans le tableau suivant :

Table IV.1. Précisions Moyennes obtenues pour les 10 classes

Base de données	Précision
Afrique	**55.02%**
Plage	**57.06%**
Monument	**33.44%**
Bus	**82.67%**
Dinosaur	**80.53%**
Éléphant	**65.34%**
Rose	**86.33%**
Cheval	**53.72%**
Montan	**49.89%**
Food	**66.36%**
TOUS	**63.04%**

4.5. Analyse des résultats

Les performances pour 30 images sur la base de données COREL obtenues par cette méthode sont les suivants :

Le meilleur résultat des classes obtenu dans la base de données est 86.33% à partir de la méthode VGG-16 pour la classe Rose. Le pire résultat est obtenu par la classe Monument avec un pourcentage de 33.44%

On remarque sur les résultats de précisions de cette méthode sont complétement différente pour chaque classe.

On peut dire, globalement, que cette méthode est utile pour rechercher des images dans le contexte d'un système CBIR.

5. Conclusion :

Dans ce chapitre nous avons implémenté VGG-16, nous avons présenté les processus de chaque méthode, aussi nous avons présenté la base de donnes utilisé par notre système et de la répartition en base d'apprentissage et base de test, ensuite on a défini l'environnement de développement, les distance utilisées, et enfin nous avons présenté les résultats obtenus avec analyse de ces dernier.

Conclusion générale

Dans ce présent mémoire nous avons abordé un domaine très intéressant qu'est la recherche d'images par le contenu.

Au début de notre travail, nous avons présenté la base d'images et la structure générale d'un système de recherche d'images par le contenu et ses composantes essentielles. Concernant ces axes, nous avons évoqué, particulièrement, les notions d'indexation et de représentation d'images dans un système CBIR.

Le système de recherche d'images par le contenu visuel se compose essentiellement de descripteurs (attributs visuels) dans une image. Ces descripteurs sont regroupés autour de : la couleur, la texture et la forme. Nous avons aussi évoqué les notions importantes des mesures de distance et des mesures de similarité qui permettent d'estimer plus ou moins efficacement le degré de similarité entre images, dans l'espace des attributs..

Nous avons, par la suite, abordé la notion d'apprentissage profond en générale et la structure et démarche générale d'un système d'apprentissage basé Deep Learning. Dans ce cadre-là, nous avons choisi comme outil pratique d'apprentissage les CNNs (Réseaux de Neurones Convolutionnels). Le système ainsi réalisé a été testé sur une partie de la base COREL 10k, base très utilisée en matière d'évaluation des systèmes CBIR.

Le travail réalisé dans le cadre de ce mémoire nous a dévoilé l'efficacité de la méthode implémenté VGG-16 et autre variantes de l'apprentissage profond, dans le cadre des systèmes CBIR.

Nous estimons avoir eu une expérience des plus intéressantes, dans le cadre de la recherche de l'image par le contenu et des techniques d'apprentissage profond.

Bibliographie

Bibliographie :

[A.C.Bov]:A. C. Bovik, M. Clark et W. S. Geisler, Multichannel texture analysis using localized

Spatial filters; journal IEEE Trans. PAMI, vol. 12 pp. 55-73, 1990.

[Agro]: P. Agouris, J. Carswell ET A. Stefanidis, An environment for content-based image retrieval from large spatial databases", Journal de photogrammetry and remote sensing, Vol.

54, No. 4, pp. 263-272, 1999.

[A.K.JAI] : A.K.Jain et F.Farrokhnia, Unsupervised texture segmentation using Gabor filters., Pattern recognition, Vol. 24, No. 12, pp. 1167-1186, 1991.

[Barb, 1984] : D. Barba et J. Ronsin, Image segmentation using new measure of texture feature. Journal de Digital Signal Processing 84, pp. 749-753, 1984

[Benl 2013] : Slimane Benloucif, « Recherche de l'image par le contenu visuel : Une approche par combinaison de plusieurs critères », Mémoire de Licence en Informatique, université de skikda, 2013.

[Bimb]: A. Del Bimbo, Visual Information Retrieval. Morgan Kaufmann Publishers, 199

[Bouk 2018] : Boukerma Rahima «La recherche de l'image par le contenu : La pondération des motifs locaux par un algorithme à évalution différentielle», Mémoire de Master Académique Option, université skikda, 2018.

[Bur, 1995] : B. Burke Hubbard, Ondes et ondelettes. Sciences d'Avenir.Belin, 1995 Open CV.

[Burger et Burge, 2009]: W. Burger, M.J. Burge, « Principles of Digital Image Processing : Core Algorithms », 1ère édition, Springer Publishing Company, Royaume Uni, 2009.

[Cire, 2013]: Ciresan, Dan; Ueli Meier; Jonathan Masci; Luca M. Gambardella; Jurgen Schmidhuber (2011). "Flexible, High Performance Convolutional Neural Networks for Image Classification" (PDF). Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume Volume Two. 2: 1237–1242. Retrieved 17 November 2013.

[Cire, 2012]: Ciresan, Dan; Meier, Ueli; Schmidhuber, Jürgen (June 2012). Multi-column deep neural networks for image classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition. New York, NY: Institute of Electrical and Electronics Engineers (IEEE).

[Datta et al., 2008]: R. Datta, D. Joshi, J. Li, J.Z. Wang,« Image retrieval : Ideas, influences, and trends of the new age », Computer Surveys, ACM, vol. 40, no. 2, art. 5, 60 pages, 2008.

[Deng]: Y. Deng ET B. Manjunath, Unsupervised segmentation of color texture regions in images and video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23:800–810, 2001.

[Fabio] : Fabio Policarpo, The Computer Image, ACM Press. Pages 298-308. 1998

[F.pret, 1988]: F. Preteux et M. Schmitt: Boolean texture analysis and synthesis. Academic Press, J. Serra Ed., Vol. 2, pp 377-400, 1988

[Fogel, 1989]: I. Fogel et D. Sagi. Gabor, filters as texture discrimination. Bio. Cybern., Vol. 61, pp. 103-113, 1989

[Grat] : C. Gratin, J. Vitria, F. Moreso et D. Seron, Texture classification using neural networks and local granulometries. Kluwer Academic Publishers, J.Serra & P. Soille Ed., pp. 309-317,1994

[Habi]: Habibi, Aghdam, Hamed (2017-05-30). Guide to convolutional neural networks: a practical application to traffic-sign detection and classification. Heravi, Elnaz Jahani. Cham, Switzerland.

[Heus]: R. Heus, "Approches virtuelles dédiées à la technologie des puces à tissus «Tissue MicroArrays » TMA : Application à l'étude de la transformation tumorale du tissu colorectal", Thèse de doctorat, Université Joseph Fourier, 28 Septembre 2009.

[Houa, 2010] : Kamel Houari «Recherche d'images par le contenu» thèse de Doctorat, Université Mentouri de Constantine, 2010.

[Hu, 1962] : M. K. Hu, Visual pattern recognition by moments invariants. Computer methods in image analysis. Transactions on Information Theory, 8, 1962.

[Ian]: Ian Goodfellow and Yoshua Bengio and Aaron Courville (2016). Deep Learning. MIT Press. p. 326.

[J. R. Smith1996]: J. R. Smith ET S.-F: Chang, Querying by color regions using the VisualSEEk contentbased visual query system, In Intelligent Multimedia InformationRetrieval. IJCAI, pp 159-173, 1996.

[Kova et Vett] : J. Kovacevic et M. Vetterli, Wavelets et Subband Coding. Prentice Hall, 1995.

[Kriz, 2012]: A. Krizhevsky, L. Sutskever and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[Kriz]: Krizhevsky, Alex. "ImageNet Classification with Deep Convolutional Neural Networks" Retrieved 17 November 2013.

[Kova] : J. Kovacevic et M. Vetterli, Wavelets et Subband Coding. Prentice Hall, 199

[LeNet]: "Convolutional Neural Networks (LeNet) – DeepLearning 0.1 documentation". DeepLearning 0.1. LISA Lab. Retrieved 31 August 2013.

[Lai]: T. S. Lai, "CHROMA, a photographic image retrieval system", PhD thesis, School of Computing, engineering and technology, University of Sunderland, UK, 20

[Mera 2011] : Melle Merabet Nabila et Melle Mahli Meriem «Recherche de l'image par le contenu» Mémoire de master en informatique en système informatique et de connaissances (S.I.C), université Abou Bekr Belkaid tlemcen 2011.

[Mesk, 2009] : Khouloud Meskaldji «Extraction et traitement de l'information : Un prototype d'un Système de recherche d'images couleurs par le contenu» Mémoire de magister Université Mentouri de Constantine, 2009.

[Rao et Loh]: A. R. Rao et G. L. Lohse: Towards a texture naming system, Identifying relevantdimension of texture. IBM Research report, RC 19140 (83352), pp.29, 199

[Rao et Loh, 1993]: A. R. Rao et G. L. Lohse, Identifying high level features of texture perception. Computer Vision, Graphics and Image Processing, Graphic Models and Image Processing, Vol. 55, pp. 218-233, 1993

[R.M.Ha]: R.M.Haralik, K. Shanmugam et I. Dinstein, Textural features for images Classification. IEEE Transaction on System , Man, Cybernetics, 3,610-621, 1973.

[Vail] : Aditya Vailaya, Mário Figueiredo, Anil Jain et HongJiang Zhang. A: Bayesian Framework for Semantic Classification of Outdoor Vacation Images, IEEE Trans. Image Processing, Vol. 10, No. 1, pp. 157-172, 2001

[Wu]: P.Wu, B.S. Manjunath, S.Newman ET H.D. Shin, A texture descriptor for browsing and Similarity retrieval, Signal processing: Image communication, vol.16, no.1, 2, pp: 33-43, 2000.

[Yama]: Yamaguchi, Kouichi; Sakamoto, Kenji; Akabane, Toshio; Fujimoto, Yoshiji (November 1990). A Neural Network for Speaker-Independent Isolated Word Recognition. First International Conference on Spoken Language Processing (ICSLP 90). Kobe, Japan.

[Y.Run, 1999] : Y. Rubner. Perceptual metrics for image database navigation. Rapport Technique CS

TR-99-1621, Stanford University, 1999.

[Y.Rub] : Y. Rubner. Perceptual metrics for image database navigation. Rapport Technique CSTR-99-1621, Stanford University, 1999.Swain M.J., Ballard D.H.(1991), “Color indexing”. International Journal of Computer Vision, vol. 7, no. 1, pp. 11-22, 199

Sitographie

[Sito1]:https://www.researchgate.net/figure/1-Principaux-composants-dun-Systeme-de-Recherche-dImages-par-le-Contenu_fig2_281012982

[Sito 2] : http://wang.ist.psu.edu/.

[Sito 3] : http://www1.cs.columbia.edu/CAVE/research/softlib/

[Sito 4]: https://myloview.fr/

[Sito 5] : http://www1.cs.columbia.edu/CAVE/curet

[Sito 6]: http://www.vision.caltech.edu/feifeili/Datasets.htm.

[Sito 7] :http://www.googleimage.com

[Sito 8]: https://stringfixer.com/fr/HSV_color_space

[Sito 9]:http://www.opticsingenieur.org/fr/cours/OPI_fr_M07_C02/co/Contenu_07.html

[Sito 10] : https://wordpress.callac.online/

[Sito 11] : http:// www.googleimage.com

[Sito 12]: https://www.juripredis.com/fr/blog/id-19-demystifier-le-machine-learning-partie-2-les-reseaux-de-neurones-artificiels

[Sito 13]: htttp://www.googleimage.com

[Sito 14]: https://www.run.ai/guides/deep-learning-for-computer-vision?fbclid=IwAR13PgtYPgdV3ht2kx9DdNY0EGS1XeTOsbSEheKLtMCGqtwrrAtbj9MePNY