République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique



Université Dr. Yahia Farès de Médéa

Faculté des Sciences et de la Technologie

Département du Génie Electrique et de l'Informatique

Projet de fin d'étude

Pour l'obtention du diplôme

De Master en Systèmes de Télécommunications

Evaluation des Deux Techniques ACP et ADL Pour L'étude et Visualisation des Données de Hautes Dimensions.

Proposé et dirigé par:

Dr. SOLTANE Mohamed.

Présenté par :

M^{elle}. ABDELLI Samia.

Promotion 2012/2013

Remerciements

A travers

Ce modeste travail, d'abord

Louange à notre Seigneur "ALLAH" qui nous a

Doté de la Merveilleuse faculté de raisonnement. Louange à

Notre Créateur qui nous a incités à acquérir le savoir.

C'est à lui que nous adressons toute notre

Gratitude en premier

Lieu.



Nous exprimons

Notre profonde gratitude à notre encadreur

Monsieur Dr SOLTANE Mohamed pour avoir bien

Voulu nous dirigé et Conseiller afin de mener à bien ce travail. Je remercie sincèrement Dr BATEL, les membres Du jury

Pour l'honneur qu'ils nous font pour examiner

Notre travail.

Que toutes les personnes qui ont

Contribué de près ou de loin à l'achèvement

De ce présent travail soient

Remerciées.

Dédicace

Je dédie

Ce modeste travail

En particulier aux êtres les plus chers à

Mon cœur, Mes parents qui m'ont donnés le courage

Et le savoir-faire, plus spécialement mon père

Qui me pousse vers le savoir et la

Connaissance.

A mes sœurs, Mes frères

Et tous mes amis, et en fin à tous ceux qui

M'ont aidé durant ma vie

Estudiantine.



Résumé :

Data mining est une technologie dont le but est la valorisation de l'information et l'extraction de connaissances d'un grand nombre de données. Cette technologie qui est devenue un outil important pour améliorer les recherches dans tous les domaines, rencontre des problèmes dus à la quantité énorme de données à exploiter. Le but de ce projet est d'introduire et d'expliquer les approches classiques de reconnaissance à savoir : ACP et ADL, et surtout de présenter les résultats de l'ensemble des tests des différentes méthodes qu'on a implémenté, des tests qui ont été réalisés sur différentes bases de données et qui ont permis de conclure que lorsqu'on utilise ACP sur une base qui respecte les conditions suivantes : un éclairage uniforme pendant les deux phases d'apprentissage et de reconnaissance, et la prise de différentes poses pour chaque visage pendant la phase d'apprentissage, et à chaque fois une légère rotation, ceci donne les performances optimales.

Mots clés: la fouille de données, ACP, ADL, reconnaissance de visage.

ملخص :

تنقيب البيانات هي تكنولوجيا، الهدف منها تثمين المعلومات و استخراج المعرفة من عدد كبير من البيانات، هذه التكنولوجيا اصبحت اداة هامة لتحسين البحث في جميع الميادين نظرا للكمية الهائلة من البيانات الواجب تنقيبها ، والغرض من هذا المشروع هو تقديم وشرح النهج الاعتراف التقليدية وهي: PCA وLDA، وخاصة لعرض نتائج جميع الاختبارات من أساليب مختلفة وقمنا بتنفيذ والاختبارات التي أجريت على قواعد بيانات مختلفة ولقد وجد أنه عند استخدام اتفاق السلام الشامل قاعدة أن تتوفر فيه الشروط التالية :الإضاءة موحدة لكلا التعلم ومراحل الاعتراف، وأخذ يشكل مختلفة لكل وجه خلال مرحلة التعلم، وفي كل مرة التناوب طفيف، وهذا يعطي افضل اداء.

كلمات البحث: التنقيب عن البيانات، LDA ، PCA ، التعرف على الوجه

Abstract:

Data mining is a technology whose purpose is the development of information and knowledge extraction of a large amount of data. This technology has become an important tool to enhance research in all fields, encounter problems due to the huge amount of data to be exploited.The goal of this project is to introduce and to explain the traditional FR approaches such as: PCA, LDA, and especially to present all the results of the tests of the various methods which we have implemented, tests which were performed on various databases. The results indicate that, when we use PCA on a database which satisfies the following conditions: an identical lighting condition during the enrollment and the verification situation, and the capture of several images of the face, showing slightly different views during the enrollment procedure, this gives the optimal performances.

Keywords: *data mining, PCA, LDA, face recognition.*

Table des Matières

Remerciements

Dédicace

Résume

Table des matières

Table des figures

CHAPITRE I : Introduction et Motivation.

CHAPITRE II : Filtrage et détection des contours.

II.1. Introduction ……. 05
II.2. Utilisation de l'opérateur Gradient ……. 06
2.1. Principes …….. 06
2.2. Quelques opérateurs gradient …….. 07
2.2.2. Opérateurs de gradient par masques ……. 07
2.2.2. b -Approximation de base …… 07
2.2.2. c -Opérateurs de Roberts ……. 09
2.2.2. d - Masques de Prewitt ……. 10
2.2.2. e - Masques de Sobel …….. 10
2.3. Filtrage de Canny …… 12
2.3.1. Principe ……. 12
2.3.2. Critère de bonne décision ……. 14
2.3.3. Critère de bonne localisation …… 15
2.3.4. Critère d'unicité de la réponse …….. 15
II.3. Utilisation du calcul du Laplacien …….. 16
3.1. L'opérateur Laplacien …….. *16*
3.1. Remède contre la sensibilité au bruit …….. 18
3.3. Le filtre de Marr ou Laplacien de Gaussienne ou chapeau mexicain …… 18
II.4. Conclusion ……. 21

CHAPITRE III : Les techniques d'exploitation de données (DATA MINING).

III.1. Introduction…… **22**
III.2. Définition du data mining..……. **22**
III.3. Principe et spécificité …… **23**
3.1. Stratégie du Data Mining …… **23**

3.2. Méthodologie …….. **24**
III.4. Data mining sur quels types de données ?..... **24**
III.5. Taxinomie des méthodes …….. **27**
5.1. Selon les objectifs ……... **27**
5.2. Selon le type d'apprentissage ……. **28**
5.3. Selon les types de modèles obtenus …… **30**
III.6. Les étapes du processus de data mining ……... **30**
III.7. Techniques du data mining …… **31**
III.8. Catégorisation des systèmes du data mining ……. **40**
III.9. Domaines d'application du data mining …… **41**
III.10. Conclusion ……. **46**

CHAPITRE IV : Evaluation d'algorithme ACP et ADL.

IV.1. Introduction ……. **47**
IV.2. Outils de Développements …….…….. **47**
IV.2.1. Matlab 7.6.0 (R2008a) ……... **47**
IV. 3. Filtrage et Détections Des Contours EVALUATION ……..... **47**
A. Lecture et Affichage d'image Numérique…….. **47**
B. Conversion des Couleurs……. **48**
C. Operations des Filtrages ……. **49**
IV.4. La reconnaissance de forme ……. **54**
IV.5. Analyse en Composantes Principales (ACP)……... **55**
IV.6. Analyse Discriminante Linéaire ……. **60**
IV.7. Bases de Données de Visages …… **62**
A. ORL 1994 [10]…… **62**
B. IFD 2002 [15] ……. **63**
IV.8. Analyse Expérimental des Résultats ……... **64**
IV.9. Conclusion …….. **67**
CHAPITRE V : Conclusion Générale……..... 68

Liste des figures :

Figure [II.1] : Principe de la détection de contour par gradient ou laplacien.	05
Figure [II.2] : Synoptique d'une détection de contours par gradient.	07
Figure [II.3] : Exemple d'application de l'approximation de base.	09
Figure [II.4] : Exemple d'application de l'opérateur de Roberts.	10
Figure [II.5] : Exemple d'application de l'opérateur de Sobel.	11
Figure [II.6] : Principe de critère de bonne décision.	14
Figure [II.7] : Principe du critère de bonne localisation.	15
Figure [II.8] : Principe du critère d'unicité de la réponse.	16
Figure [II.9] : Exemple d'obtention des contours par utilisation du Laplacien suivi d'un seuillage.	18
Figure [II.10] : Allure de la fonction Laplacien de Gaussienne pour deux valeurs de σ.	19
Figure [II.11] : Application du filtre LOG à la détection de contours.	20
Figure [III.1] : Techniques du data mining.	22
Figure [III.2] : Architecture du Data mining.	23
Figure [III.3]: Processus de data mining.	31
Figure [III.4] : Réseau de neurones.	32
Figure [IV.1] : Affichage histogramme d'image .	48
Figure [IV.2]: Conversion des couleurs.	49
Figure [IV.3]: Operateur de filtrage Gaussien.	50
Figure [IV.4]: Filtrage pour les déférents alphas.	51
Figure [IV.5]: Détection des contours.	51
Figure [IV.6]: Détection des bruits.	53
Figure [IV.7]: Application de filtre LOG à la détection de contours.	53
Figure [IV.8] : Détection des contours selon les opérateurs Log, Sobel et Prewitt.	54
Figure [IV.9] : Interprétation géométrique.	58
Figure [IV.10] : Les classifications de PCA.	60
Figure [IV.11] : Images d'une personne dans la base IFD.	63
Figure [IV.12] : Le pourcentage de reconnaissance pour les algorithmes PCA et LDA.	64
Figure [IV.13]:Temps de traitement pour les algorithmes de PCA et LDA sur la base ATT et ITT.	65
Figure [IV.14] : Temps de test pour les algorithmes PCA et LDA sur la base ATT et IFD.	66

Introduction et Motivations.

En raison de l'augmentation constante du volume d'informations accessible électroniquement, la conception et la mise en œuvre d'outils efficaces, permettant notamment à l'utilisateur de n'avoir accès qu'à l'information qu'il juge pertinente, devient une nécessité absolue. Comme la plupart de ces outils sont destinés à être utilisés dans un cadre professionnel, les exigences de fiabilité et de convivialité sont très importantes ; les problèmes à résoudre pour satisfaire ces exigences sont nombreux et difficiles. L'accès à l'information pertinente peut être réalisée en fournissant à un utilisateur des documents pertinents, ou en lui proposant des passages de documents pertinents (ou des réponses à des questions). Le premier cas relève du domaine de la recherche de textes, le second du domaine de l'extraction d'informations [21].

L'exploration de données, aussi connue sous le nom fouille de données, datamining (forage de données) ou encore extraction de connaissances (ECD en français, KDD en Anglais), a pour objet l'extraction d'un savoir 'utile' ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. *Datamining* étant un processus d'extraction de connaissances valides et exploitables à partir de volumes de données même de tailles considérables, il a vocation à être utilisé dans un environnement professionnel et se distingue de l'analyse de données et des techniques de la statistique par les points suivants :

- Contrairement aux méthodes statistiques, *Datamining* ne nécessite jamais que l'on établisse une hypothèse de départ qu'il s'agira de vérifier. C'est à partir des données elles-mêmes que se dégageront les corrélations intéressantes, l'algorithme n'étant là que pour les découvrir et non pour les établir. *Datamining* se situe à la croisée des statistiques, de l'intelligence artificielle et des bases de données.
- Les connaissances extraites par *Datamining* ont vocation à être intégrées dans le schéma organisationnel de l'entreprise ou de l'entité considérée. *Datamining* impose donc d'être capable d'utiliser de manière opérationnelle les résultats des analyses effectuées, souvent dans des délais très courts. Le processus d'analyse doit permettre à l'organisation une réactivité très importante.

- Les données traitées sont issues des systèmes de stockage en place dans l'organisation et sont ainsi hétérogènes, multiples, plus ou moins structurées. Pour résumer ce sont des données dont la raison d'être n'est a priori pas l'analyse. Cela impose de disposer de systèmes performants de préparation ou de manipulation de données [8].

Datamining se propose donc de transformer en information, en connaissance, de grands volumes de données qui peuvent être stockées de manières diverses dans des bases de données relationnelles ou dans un (ou plusieurs) entrepôt de données (*datawarehouse*), mais qui peuvent aussi être récupérées de sources riches plus ou moins structurées comme Internet, ou encore en temps réel. Lorsque la source n'est pas directement un entrepôt de données, il s'agira très souvent de construire une base de données ou *datamart* dédié à l'analyse et aux analystes. Cela suppose bien entendu d'avoir à sa disposition une palette d'outils de gestion de données (*data management*).

« La nécessité est la mère de l'invention. — Platon »

Datamining a eu une grande attention ces derniers temps dans l'industrie de l'information, ça est due principalement à la disponibilité de grandes quantités de données et au besoin de les transformés en connaissances utiles surtout dans les applications concernant l'analyse du marché, la détection de fraudes, la conservation de client, le contrôle de productions et l'exploration scientifique. *Datamining* peut être vue comme l'évolution naturelle de la technologie d'information. Depuis les années 60s, la technologie des informations et des Bases de Données a systématiquement évoluée des systèmes de gestion de fichiers aux systèmes de Bases de données plus puissant et plus sophistiqués. La recherche et le développement dans le domaine de systèmes de Bases de Données depuis les années 70s a menée au développement des systèmes de Bases de données relationnels, des outils de modélisation des données et des techniques d'indexation et d'organisation des données.

Depuis le milieu des années 80s, la technologie des bases de données a été caractérisée par l'adoption de la technologie relationnelle et la croissance des recherches et des activités de développement d'un nouveau et puissant système de bases de données [ ].

Ceux-ci favorisent des modèles avancés de données comme le modèle relationnel étendu, orienté-objet, relationnel-objet et des systèmes de Bases de données déductifs. Des issues de distribution, de diversification et de partage de données ont été étudiées intensivement. Des systèmes de BD hétérogènes et des systèmes d'informations basés sur

l'internet comme le World Wide Web (www) a aussi émergés à jouer un rôle vital dans l'industrie de l'information.

La progression de la technologie du matériel informatique pendant les trois décennies passées nous a amené à la puissance accrue des ordinateurs, des équipements de collection de données et du stockage du media. Cette technologie a donné un grand coup on puce à l'industrie de bases de données et de l'information et a permet la gestion des transactions, la recherche d'informations et l'analyse de données de très grandes bases de données.

Les données peuvent maintenant être stockées dans des différentes bases de données. L'une de Bases de données la plus récente est le *Data Warehouse,* qui regroupe une multitude de sources de données hétérogènes, organisées sous un unie schéma et un seul emplacement pour faciliter la gestion. La technologie du Data *Warehousing* incluse l'intégration de données et le XOLAP (*On-Line Analytical Processing)* qui est un ensemble de techniques d'analyse avec des fonctionnalités tels que la consolidation et l'agrégation.

L'abondance de données couplées avec le besoin de bons outils d'analyse de données ont été décrits par "*Riche en données mais Pauvre en informations*". La croissance rapide et l'énorme quantité de données collectées et stockées dans de grandes et nombreuses bases de données ont excédé notre habilité humaine de vouloir comprendre sans utiliser aucuns outils. Et comme résultat à tout ça, les grandes bases de données collectant de grandes quantités de données sont devenues des tombes de données (les données archivées sont rarement revisitées). Par conséquence, des décisions importantes sont souvent prises par l'intuition des décideurs et non pas par les données riches en informations bases de données, simplement par ce connaissances précieuses embarquées dans les grandes quantités de données [4].

Les technologies des systèmes experts actuels s'appuient typiquement sur les utilisateurs ou les experts d'un domaine pour faire entrer des connaissances dans des bases de connaissances manuellement. Malheureusement, cette procédure est des erreurs et elle prend du temps. Les outils du *Datamining* qui exécutent l'analyse des données peuvent découvrir d'important modèles de données stratégies du business, aux bases de connaissances et a médicales. L'élargissement de l'écart entre données et informations fait l'appel au développement systématique des outils des données en pépites d'or de connaissances [4].

L'objectif de ce travail est évaluation des deux technique PCA et LDA pour l'étude et visualisation des données de hautes dimensions. Pour ce fait notre mémoire nous commençons par un chapitre qui est justement de permettre au lecteur de connaître plusieurs méthodes et de savoir agir sur les paramètres de ces méthodes. Dans ce but, la première partie décrit des techniques de calcul du Gradient (détecteurs par masques de Roberts, Prewitt,

Sobel). La seconde partie aborde de filtrage de Canny, les calculs Laplacien (filtre laplacien, filtre « chapeau mexicain », filtre différence de gaussiennes).

Le troisième chapitre est introduit la notion de datamining par donner quelques définitions de ce terme ainsi les principes et les spécificités, ensuite nous avons illustré les différents types de données qu'un système de datamining peut s'appliquer, après ça nous allons expliquer toutes les taches du datamining et que chacune d'elles représente une façon d'extraction de connaissances. Les étapes du processus de datamining ont été aussi expliquées tout comme les techniques et les algorithmes utilisées dans le domaine de datamining et la catégorisation des systèmes de datamining, et pour conclure ce chapitre nous avons abordées avec un peu de détails les différents domaines d'application de datamining.

Le quatrième chapitre contient une évaluation d'algorithme ACP et ADL sur la base de données ATT et IDF.

Nous finissons notre mémoire par une conclusion générale qui contient les perspectives possibles qui constituent une suite de recherche pour notre projet.

Filtrage et Détection des Contours.

II.1.Introduction :

La mise en évidence des points représentant les contours d'objets dans une image peut servir à reconnaître des objets présents dans une scène, à différencier des zones de l'image, à faire de la segmentation d'images, à extraire une information réduite souvent pertinente pour caractériser l'image.

Un contour se matérialise par une rupture d'intensité dans l'image suivant une direction donnée. Plusieurs méthodes existent pour détecter cette rupture, les unes plus ou moins complexes, les autres plus ou moins gourmandes en calculs. Dans la plupart des cas et en particulier pour ceux présentées ici, la même méthodologie est employée. Elle s'applique en deux étapes : la première permet de localiser les contours à partir d'un calcul de Gradient ou de Laplacien dans des directions privilégiées tout en quantifiant l'importance du contour. La seconde étape va permettre d'isoler les contours du reste de l'image à partir d'un seuillage judicieux. Plusieurs méthodes permettent de déterminer le Gradient ou le Laplacien d'une image. Il en est de même des techniques de seuillage. Ces deux étapes sont indépendantes, il existe donc un grand nombre de combinaisons calcul de Gradient-opération de seuillage conduisant à la mise en évidence des contours. Le rôle de l'ingénieur est souvent de choisir les opérateurs les mieux à même de satisfaire les exigences associées à l'image à traiter [1] .

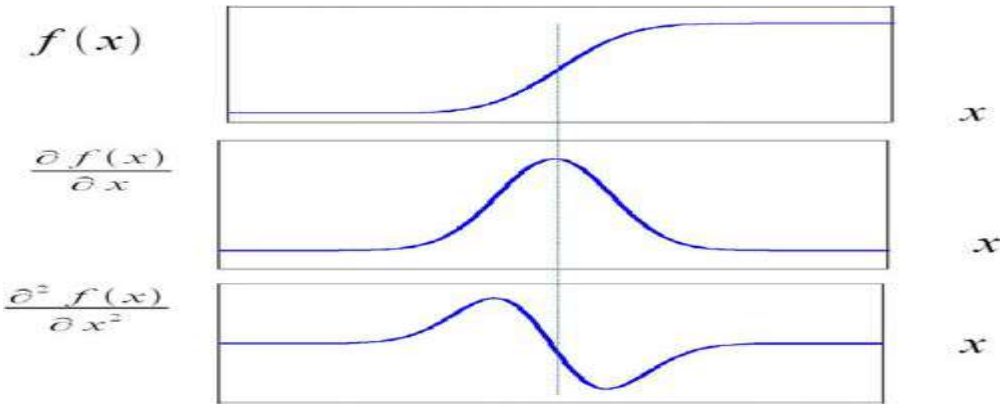


Figure [II.1] : Principe de la détection de contour par gradient ou laplacien [1].

L'objectif de ce chapitre est justement de permettre au lecteur de connaître plusieurs méthodes et de savoir agir sur les paramètres de ces méthodes. Dans ce but, la première partie décrit des techniques de calcul du Gradient (détecteurs par masques de Roberts, Prewitt, Sobel, , filtrage de Canny). La seconde partie aborde les calculs de Laplacien (filtre laplacien, filtre « chapeau mexicain », filtre différence de gaussiennes).

II.2.Utilisation de l'opérateur Gradient :

2.1. Principes :

En considérant l'image dans un repère orthogonal (O_{xy}) tel que (O_x) désigne l'axe horizontal et(O_y) l'axe vertical, le Gradient de l'image (ou plutôt de la luminance f) en tout point ou pixel de coordonnées(x, y) est désigné par :

$$\overrightarrow{\mathrm{Grad}}\,\mathrm{f} = \vec{\nabla}\mathrm{f} = \begin{pmatrix} \frac{\partial \mathrm{f}}{\partial \mathrm{x}} \\ \frac{\partial \mathrm{f}}{\partial \mathrm{y}} \end{pmatrix} \qquad [\mathrm{II.1}]$$

Le module du gradient permet de quantifier l'importance du contour mis en évidence, c'est-à-dire l'amplitude du saut d'intensité relevé dans l'image :

$$\left\|\overrightarrow{\nabla \mathrm{f}}\right\| = \sqrt{\left(\frac{\partial \mathrm{f}}{\partial \mathrm{x}}\right)^2 + \left(\frac{\partial \mathrm{f}}{\partial \mathrm{y}}\right)^2} \qquad [\mathrm{II.2}]$$

La direction du gradient permet de déterminer l'arête présente dans l'image. En effet, la direction du gradient est orthogonale à celle du contour :

$$\alpha_0 = \arctan\left(\frac{\frac{\partial \mathrm{f}}{\partial \mathrm{y}}}{\frac{\partial \mathrm{f}}{\partial \mathrm{x}}}\right) \qquad [\mathrm{II.3}]$$

Le principe de la détection de contours par l'utilisation du gradient consiste à calculer d'abord le gradient de l'image dans deux directions orthogonales puis le module du gradient. Il s'agira ensuite d'effectuer une sélection des contours les plus marqués, c'est-à-dire les points de plus fort contraste par un seuillage adéquat, les directions des contours étant orthogonales à la

direction α_0 déterminée en tout pixel de l'image (voir ci-dessous). Les contours sont finalement représentés par des pixels blancs sur fond noir [1].

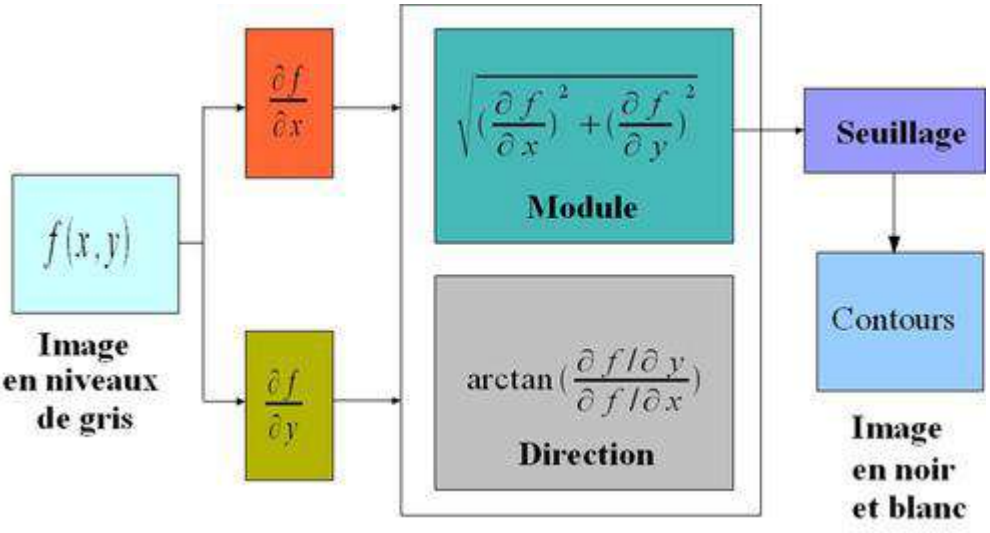


Figure [II.2] : Synoptique d'une détection de contours par gradient [1].

2.2. Quelques Opérateurs Gradient :

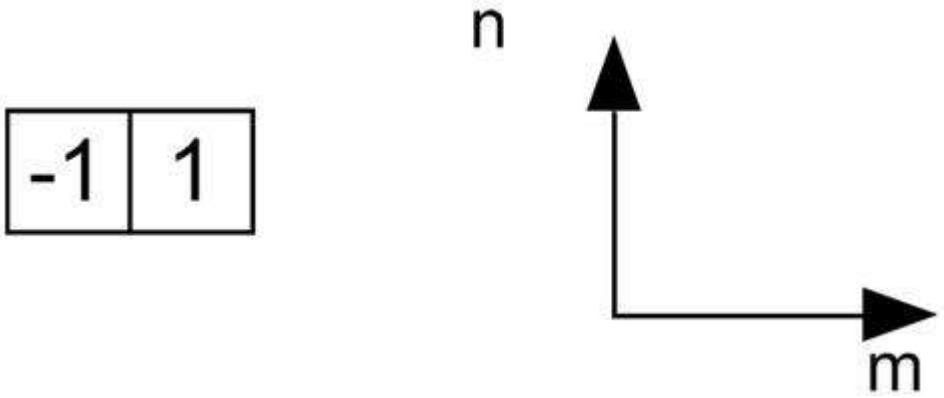
Ces opérateurs sont à considérer comme des filtres qui vont être corrélés à l'image. Les réponses impulsionnelles de ces filtres peuvent se présenter sous la forme de fonctions analytiques souvent d'une seule variable ou bien sous la forme de masques bidimensionnels. Dans les deux cas le filtrage a lieu en deux étapes : un filtrage suivant les lignes de l'image puis suivant les colonnes dans le cas d'une expression monodimensionnelle de la réponse impulsionnelle du filtre, une corrélation bidimensionnelle de l'image avec deux masques modélisant deux contours dans des directions orthogonales dans l'autre cas.

2.2.1. Opérateurs de Gradient par Masques :

Pour chaque opérateur, deux masques sont utilisés de façon à déterminer le gradient de l'image dans deux directions orthogonales.

2.2.1. a -Approximation de Base :

Le masque le plus intuitif à mettre en œuvre est un masque à deux éléments :



L'origine du masque est le point -1.

Dans le repère ci-dessus, la réponse impulsionnelle $h(m,n)$du masque est définie par :

$$h(0,0) = -1 \qquad h(1,0) = 1$$

La corrélation de ce masque avec une image de luminance $f(i,j)$ s'écrit :

$$\sum_{m=0}^{1}\sum_{n=0} h(m,n)f(m+i,n+j) = h(0,0)f(i,j) + h(1,0)f(i+1,j)$$

$$= f(i+1,j) - f(i,j) \qquad [II.4]$$

Il s'agit bien d'un calcul de gradient suivant l'axe horizontal. En faisant subir une rotation de $\frac{\Pi}{2}$ au premier masque, il apparaît le filtre suivant dont l'origine est le point 1 :

1
-1

La réponse impulsionnelle est telle que :

$$h(0,0) = 1 \qquad h(0,-1) = -1$$

La corrélation de ce masque avec l'image *f* (*i*, *j*) permet bien d'implanter un gradient dans la direction verticale :

$$\sum_{m=0}^{0}\sum_{n=-1} h(m,n)f(m+i,n+j) = h(0,0)f(i,j) + h(0,-1)f(i,j-1)$$

$$= f(i,j) - f(i,j+1) \qquad [II.5]$$

La figure suivante propose une illustration de l'application de ces masques.

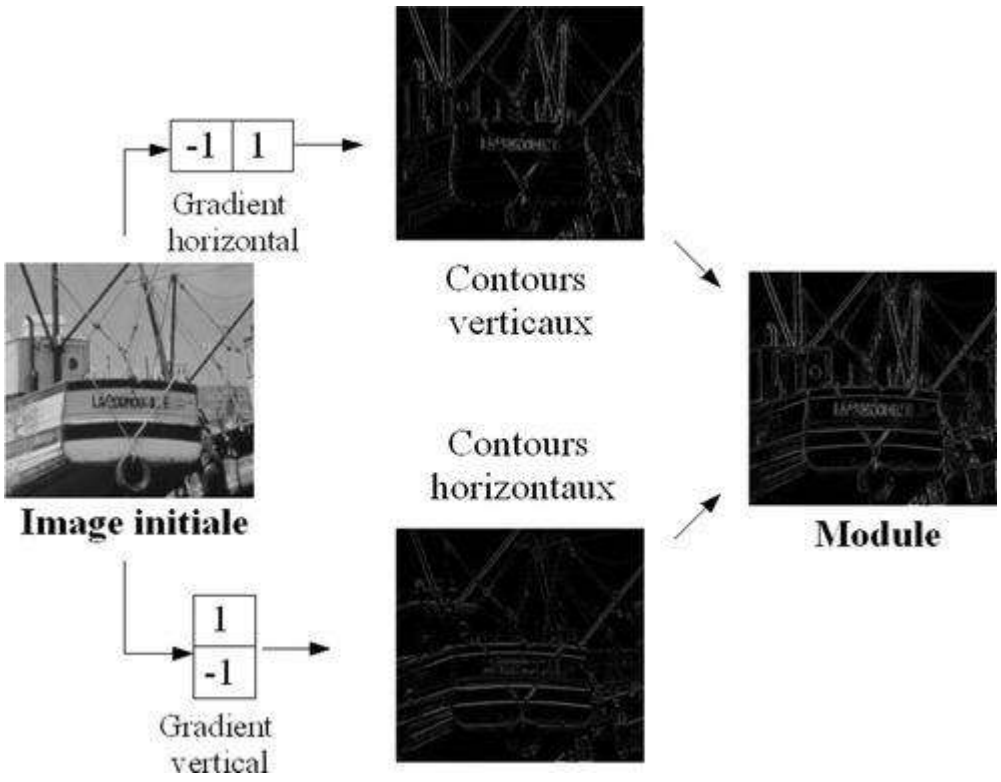


Figure [II.3] : Exemple d'application de l'approximation de base [1].

2.2.2. b - Opérateurs de Roberts [11]:

L'operateur Roberts a été créé en 1965, il approxime des dérivées directionnelles suivant les axes orientés à 45 degrés à l'aides d'une convolution de l'image f(i,j) avec les masques 2*2 suivant[2] :

$$h1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad h2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Cet opérateur présente une forte sensibilité au bruit en raison de la taille des masques (nécessité de faire un préfiltrage) [2] [11].

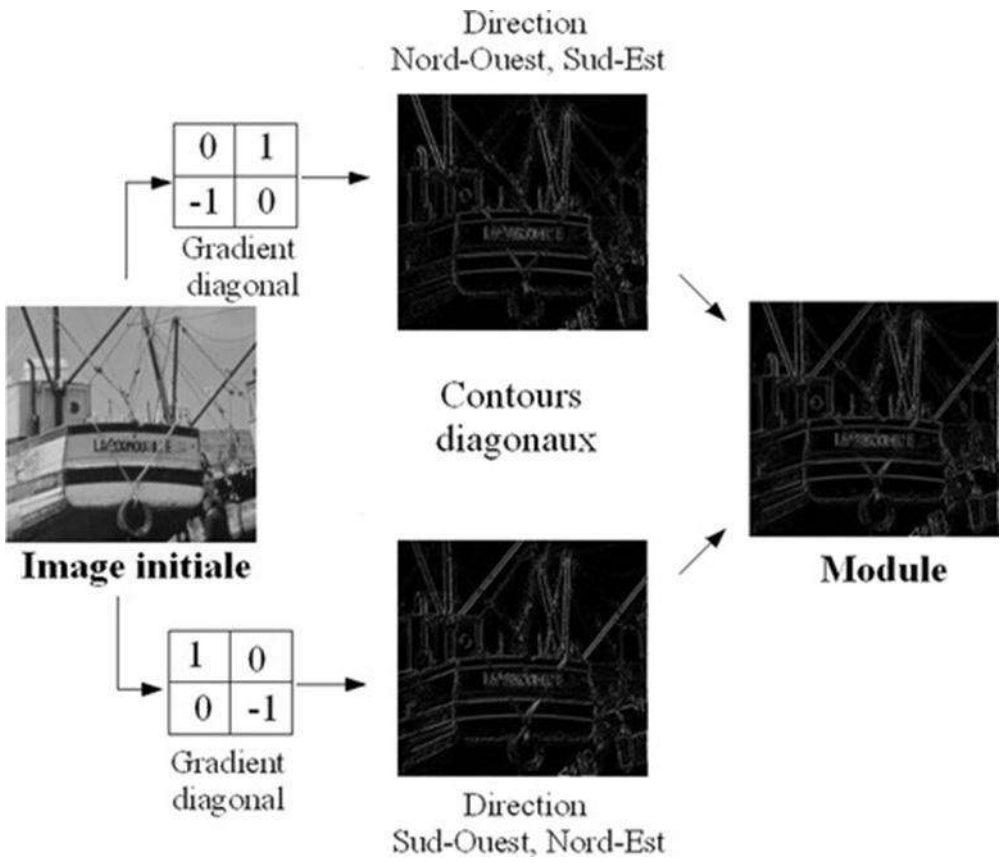


Figure [II.4] : Exemple d'application de l'opérateur de Roberts [1].

2.2.2. c - Masques de Prewitt (1970) [11] :

$$h1 = 1/3\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \qquad h2 = 1/3\begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

- La convolution de l'image par les masques ci-dessus correspond au calcul des dérivées de l'image pré filtrée par le filtre séparable:
  - h1 : dérivation horizontale
  - h2 : dérivation verticale

2.2.2. d - Masques de Sobel (1972) [11] :

$$h1 = 1/4\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \qquad h2 = 1/4\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

La dérivation accentue le bruit de l'image, c'est-à-dire les pixels de valeur parasite et de répartition aléatoire. Les opérateurs de Sobel, qui effectuent une moyenne locale sur trois pixels en largeur, sont moins sensibles au bruit :

Cet opérateur est très populaire. Il correspond à la convolution de l'image par : $[1 \quad 2 \quad 1] * [-1 \quad 0 \quad -1]$

Ce sont des masques directionnels également (Nord Est)

$$\begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix}$$ Dérivation oblique

Le majeur inconvénient de ces masques réside dans leur forte sensibilité au bruit du fait de l'implantation de la dérivation qui se traduit par un filtrage passe-haut. D'autres masques ont ainsi été proposés afin de rendre le filtrage moins sensible au bruit [11] [2].

La figure suivante propose une illustration de l'application des masques de Sobel.

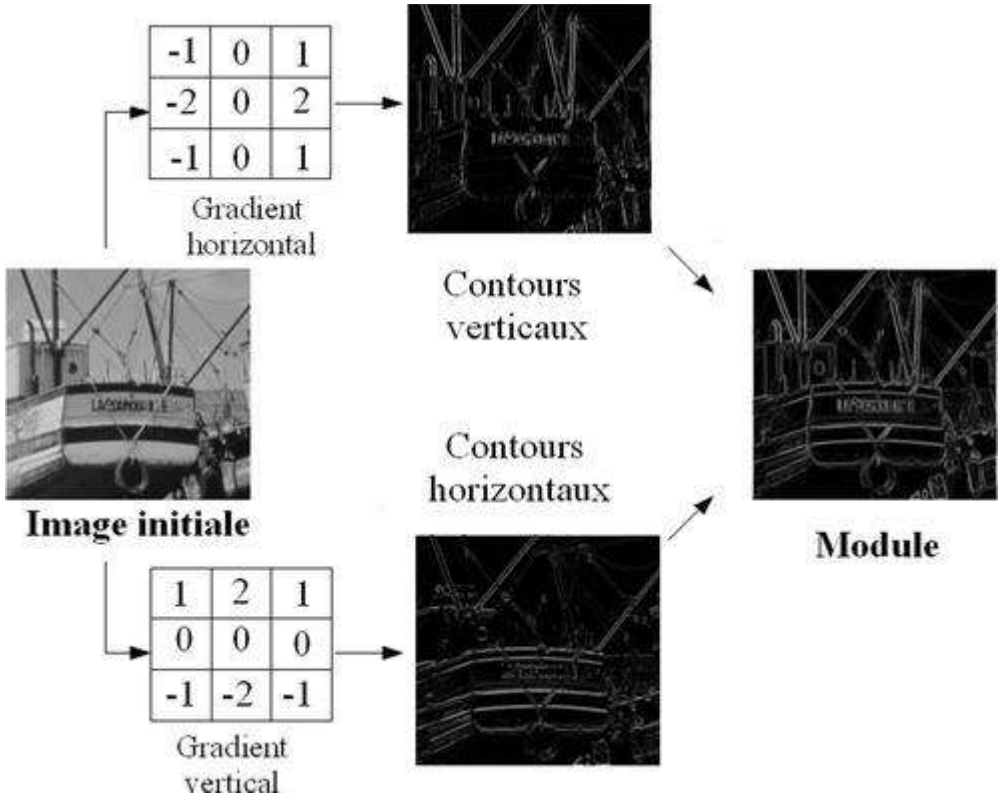


Figure [II.5] : Exemple d'application de l'opérateur de Sobel [1].

Il existe bien sûr beaucoup d'autres masques utilisés pour déterminer le gradient d'une image (Frei-Chen, ...). Le principal intérêt de ces masques est leur facilité de mise en œuvre ainsi que la rapidité de leur traitement. Leur inconvénient est leur grande sensibilité au bruit. De plus les contours obtenus sont souvent assez larges. Le filtre de Sobel est le plus utilisé dans les applications industrielles nécessitant des contraintes temps-réel [1].

2.3. Filtrage de Canny :

Canny a proposé en 1983 une étude théorique de la détection de contour .son étude s'est limite au cas de la dimension 1, s'est à dire la détection des variables dans un signal bruité [2].

2.3.1-Principe :

Les approches précédentes par masques étaient basées sur une modélisation assez simple d'un contour idéal. Pour le calcul du filtre de Canny, une approche analytique plus élaborée est employée. Il s'agit d'une technique de filtrage optimal. Canny a en effet cherché à déterminer de façon analytique en 1986 un filtre à partir de trois critères :

- un critère de bonne détection garantissant une réponse forte en sortie du filtre même en présence de faibles contours sur l'image d'entrée,
- un critère de bonne localisation du contour,
- un critère d'unicité de la réponse permettant d'assurer une seule détection pour un contour et ainsi d'éviter les effets de rebond.

Canny définit ces trois critères de façon mathématique. L'optimisation des trois critères proposés permet de définir le filtre linéaire optimal pour la détection d'une marche d'escalier sous l'hypothèse d'un bruit additif indépendant du signal. Il s'agit de trouver la réponse impulsionnelle $h(x)$ du filtre optimal qui permet d'obtenir une valeur maximum en sortie lorsqu'un contour est présenté en entrée. Canny souhaite un filtre à réponse impulsionnelle finie [1] [11].

Le modèle de contour utilisé est une marche d'escalier :

$$e(x) = Ay(x) + n(x) \qquad [II.6]$$

Où *A* représenté l'amplitude du saut d'intensité, *n(x)* un bruit blanc additif indépendant de l'arête et *Y(x)* la fonction d'heaviside définie telle que :

$$y(x) = \begin{cases} 1 \text{ si } x \geq 0 \\ 0 \text{ sinon} \end{cases} \quad [II.7]$$

La définition de l'arête est telle que le saut d'intensité se situe en *x=0*. Les hypothèses concernant le bruit additif sont :

$$\begin{cases} E[n(x) = 0] \\ E[n(x)n(y)] = \sigma_0^2 \sigma(x-y) \end{cases} \quad [II.8]$$

Où E [n(x)=0] désigne l'espérance mathématique.

La moyenne des échantillons est nulle et les échantillons sont décarrelés. La sortie *g*(*x*) s'écrit comme la convolution de l'entrée e (x) par la réponse impulsionnelle du filtre recherché h(x) :

$$\begin{aligned} g(x) &= e(x) * h(x) \\ &= \int_{-\infty}^{+\infty} h(y)e(x-y)dy \\ &= \int_{-\infty}^{+\infty} Ah(y)y(x-y)dy \\ &+ \int_{-\infty}^{+\infty} h(y)n(x-y)dy \end{aligned} \quad [II.9]$$

Or:

$$y(x) = \begin{cases} 1 \; si \; y \leq x \\ 0 \; sinon \end{cases}$$

D'où :

$$g(x) = \int_{-\infty}^{+\infty} Ah(y)dy + \int_{-\infty}^{+\infty} h(y)n(x-y)dy \quad [II.10]$$

Le premier terme correspond au signal utile, le second au bruit. Du premier terme, on déduit :

$$g'(x) \approx h(x)$$

Le filtre est donc un dérivateur.

Or pour mettre en évidence le contour, la sortie $g(x)$ doit être maximale à l'endroit du contour, c'est-à-dire en $x=0$. Cette sortie sera donc paire et on en déduit que la réponse impulsionnelle $h(x)$ du filtre optimal est impaire [1].

2.3.2-Critère de Bonne Décision :

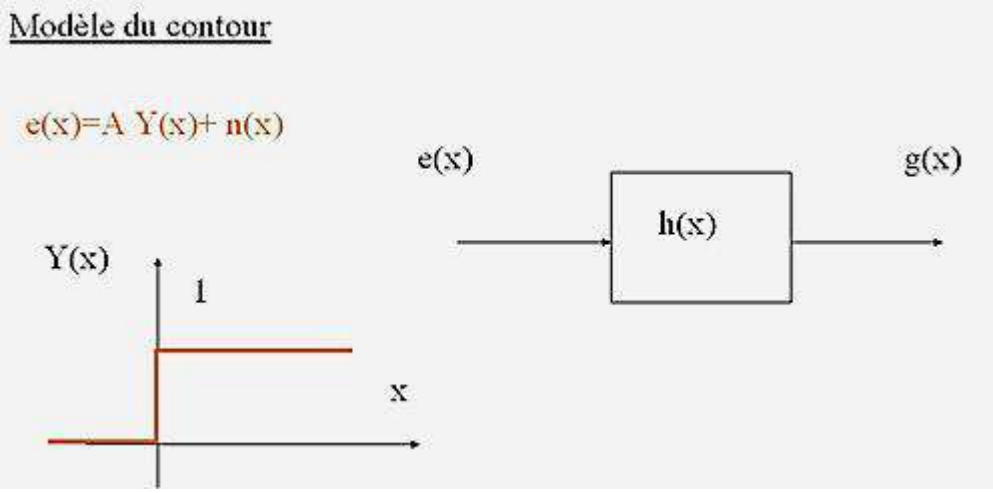


Figure [II.6] : Principe de critère de bonne décision [1].

Le premier critère choisi par Canny est celui de bonne décision qui consiste à maximiser le rapport signal à bruit (RSB) en sortie du filtre même en cas de contour faible. Il est défini à l'endroit du contour en $x=0$ comme le rapport du maximum de la réponse due au signal sur la valeur efficace du bruit :

$$RSB_{X=0} \frac{\text{amplitude du signal}}{\sqrt{\text{variance du bruit}}} \qquad [II.11]$$

Canny obtient le critère suivant à maximiser (« Critère de bonne décision ») :

$$\frac{A}{\sigma_0}\sum(h)$$

Avec :

$$\sum(\mathrm{h}) = \frac{\int_{-\infty}^{0} \mathrm{h}(0)\mathrm{dy}}{\sqrt{\int_{-\infty}^{+\infty} \mathrm{h}^2\,(\mathrm{y})\mathrm{dy}}} \qquad [\mathrm{II.12}]$$

2.3.3-Critère de Bonne Localisation :

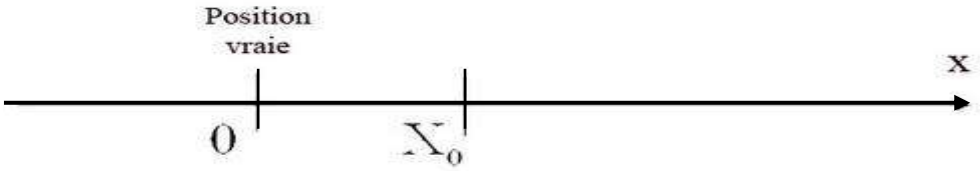


Figure [II.7] : Principe du critère de bonne localisation [1].

En présence d'un contour noyé dans du bruit, il est probable que la sortie du filtre soit maximum en $x=X_0$ et non pas en $x=0$. Soit *X0* la variable aléatoire correspondant à la distance entre le maximum de la réponse $g(x)$ et la position réelle de la transition ($x=0$). Le second critère à maximiser proposé par Canny est l'inverse de l'écart-type de la variable aléatoire X_0 :

$$\frac{1}{\sqrt{\mathrm{E}[\mathrm{X}_0^2]}}$$

$$\frac{\mathrm{A}}{\sigma_0}\Lambda(\mathrm{h})$$

Canny cherche donc bien à minimiser la variance de l'écart entre la position du maximum en sortie de filtrage et la position réelle du contour [1].

Avec :

$$\Lambda\,(\mathrm{h}) = \frac{|\mathrm{h}'(0)|}{\sqrt{\int_{-\infty}^{+\infty} \mathrm{h}'^2(\mathrm{y})\mathrm{dy}}} \qquad [\mathrm{II.13}]$$

2.3.4-Critère d'unicité de la Réponse :

Canny cherche à obtenir une faible multiplicité des maxima dus au bruit en décidant de maximiser la distance entre deux maxima de la sortie.

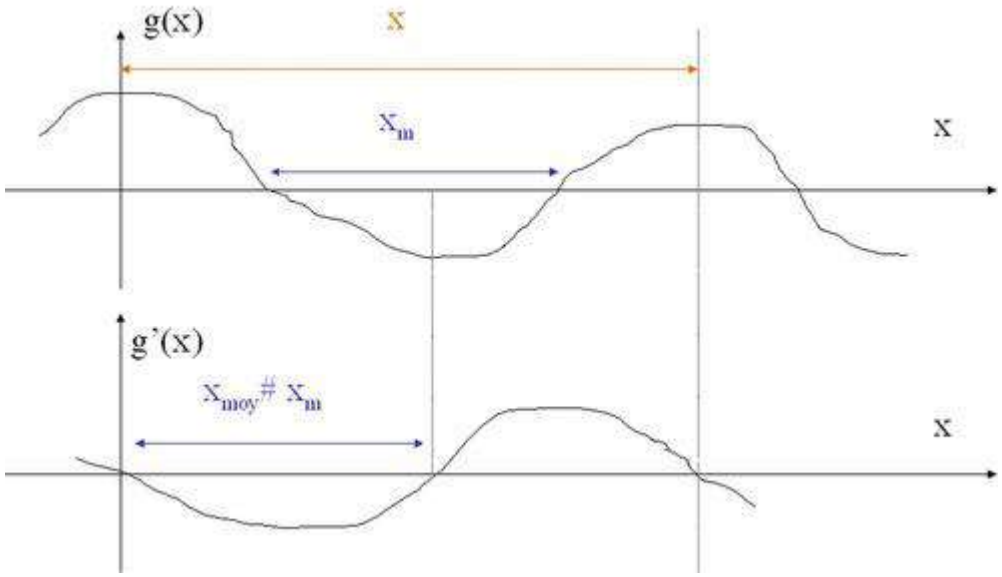


Figure [II.8] : Principe du critère d'unicité de la réponse [1].

La distance à maximiser est donnée par x sur la sortie g(x), ce qui équivaut à $2x_{moy}$ sur la sortie g'(x).

Le troisième critère s'écrit :

$$x = 2\pi\left(\frac{\int_{-\infty}^{+\infty} h'^2(y)dy}{\int_{-\infty}^{+\infty} h''^2(y)}\right)^{1/2} \qquad [II.14]$$

II.3.Utilisation du Calcul du Laplacien :

3.1. Opérateur Laplacien :

Les opérateurs de gradient vus précédemment exploitent le fait qu'un contour dans une image correspond au maximum du gradient dans la direction orthogonale au contour. Or le passage par zéro de la dérivée seconde d'une rupture d'intensité permet également de mettre en évidence le contour. La dérivée seconde est donc déterminée par le calcul du Laplacien :

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) \qquad [II.15]$$

Qui peut s'écrire :

$$\nabla^2 f = \nabla_x(f(x+1,y) - f(x,y) + \nabla_y\big(f(x,y+1) - f(x,y)\big) \qquad [II.16]$$

Or on peut définir :

$$\nabla_x\big(f(x+1,y)-f(x,y)\big)=f(x+1,y)-f(x,y)-f(x,y)-f(x-1,y))$$

$$\nabla_x(f(x+1,y)-f(x,y))=f(x+1,y)-f(x-1,y)-2f(x,y)) \qquad [II.17]$$

Et :

$$\nabla_y\big(f(x,y+1)-f(x,y)\big)=f(x,y+1)-f(x,y)-f(x,y)-f(x,y-1))$$

$$\nabla_y(f,y+1)-f(x,y))=f(x,y+1)-f(x,y-1)-2f(x,y) \qquad [II.18]$$

Ainsi le Laplacien peut s'écrire :

$$\nabla^2 f=f(x+1,y)+f(x-1,y)+f(x,y+1)+f(x,y-1)-4f(x,y) \qquad [II.19]$$

Cette opération de calcul de Laplacien peut alors être appliquée à une image par l'intermédiaire d'un filtrage avec le masque 3*3 suivant :

0	1	0
1	-4	1
0	1	0

D'autres masques peuvent être utilisés

0	-1	0
-1	4	-1
0	-1	0

-1	-1	-1
-1	-8	-1
-1	-1	-1

1	-2	1
-2	4	-2
1	-2	1

Après le filtrage de l'image au moyen d'un de ces filtres, il s'agit de détecter les passages par zéro en conservant uniquement les passages les plus marqués. En effet, la technique est particulièrement sensible au bruit en raison de la double dérivation. Il s'agit donc de ne pas considérer le bruit, qui peut très bien se traduire par des oscillations autour de zéro, comme un contour. C'est le rôle du seuil S qui va être utilisé dans cette approche pour ne prendre en compte que les passages par zéro d'amplitude relativement élevée correspondant à des vrais contours de l'image [1] [9].

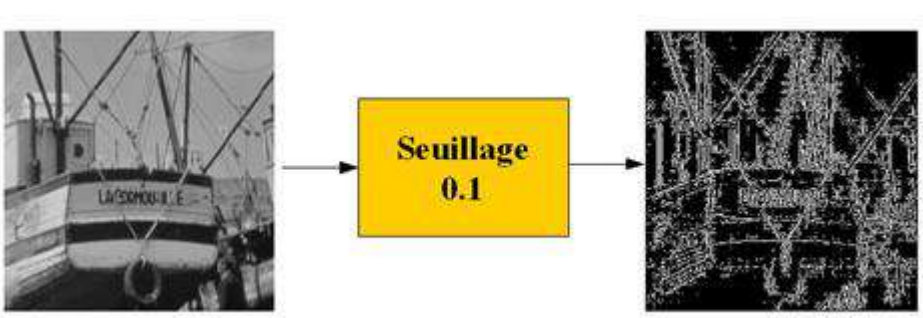


Figure [II.9] : Exemple d'obtention des contours par utilisation du Laplacien suivi d'un seuillage [1].

Remarque :

Une animation dans l'étude de cas permet de paramétrer le seuil après le calcul du Laplacien.

3.1. Remède Contre la Sensibilité au Bruit :

De même que lors de l'utilisation de l'opérateur gradient, l'idée consiste d'abord à filtrer passe-bas l'image avant d'appliquer l'opérateur Laplacien. Mathématiquement cela revient à convoluer l'image initiale inchangée $f(x,y)$ avec la dérivée seconde de la réponse impulsionnelle $h(x,y)$ du filtre passe-bas comme le montre l'écriture suivante dans le cas monodimensionnel :

$$\frac{\partial^2}{\partial x^2}\big(f(x,y) * h(x,y)\big) = \frac{\partial^2 f(x,y)}{\partial x^2} * h(x,y) = f(x,y) * \frac{\partial^2 g(x,y)}{\partial x^2} \qquad [II.20]$$

3.3. Le filtre de Marr ou Laplacien de Gaussienne ou Chapeau Mexicain :

Le remède contre la sensibilité au bruit de la double dérivation est à la base des travaux de Marr qui ont débouché sur l'utilisation d'un filtre gaussien en guise de filtre passe-bas pour $h(x,y)$. Le choix de ce filtre a été établi en raison du principe d'Heisenberg qui stipule que l'on ne peut pas améliorer en même temps la résolution spatiale Δx et la résolution fréquentielle Δf :

$$\Delta_x \Delta f \geq \frac{1}{4\pi}$$

Dans le cas d'une détection d'un contour, celui-ci étant localisé spatialement et en fréquence, il est clair que les intervalles Δx et Δf doivent être petits. Détecter une rupture d'intensité

implique une analyse des pixels voisins mais le voisinage ne doit pas être trop étendu. Filtrer l'image a pour conséquence de réduire les variations d'intensité, ce qui doit être raisonnable. Le filtre doit en effet avoir une largeur spectrale limitée. Or c'est la fonction gaussienne qui permet d'obtenir l'égalité dans le principe d'Heisenberg, d'où ce choix qui se traduit par :

$$h(x,y) = \frac{1}{2\pi\,\partial^2} e^{\frac{-(x^2+y^2)}{2\,\partial^2}} \qquad [II.21]$$

La dérivée de la fonction gaussienne bi-dimensionnelle s'écrit alors :

$$h'(x,y) = -\frac{(x+y)}{2\pi\,\partial^4} e^{\frac{-(x^2+y^2)}{2\,\partial^2}} \qquad [II.22]$$

Et la dérivée seconde s'exprime :

$$h''(x,y) = -\frac{1}{\pi\,\partial^4}\left(1 - \frac{x^2+y^2}{2\,\partial^2}\right) e^{-\frac{x^2+y^2}{2\,\partial^2}} \qquad [II.23]$$

L'allure de h"(x,y), réponse impulsionnelle du filtre Laplacien de Gaussienne (LOG), lui a valu l'appellation « chapeau mexicain ». La mise en œuvre d'une détection de contours avec le filtre LOG nécessite les étapes suivantes :

- Convolution de l'image avec le filtre LOG (filtre 2D) ;
- Détection des passages par zéro de l'image résultante ;
- Seuillage de l'image afin de ne considérer que les passages par zéro d'amplitude suffisante [1].

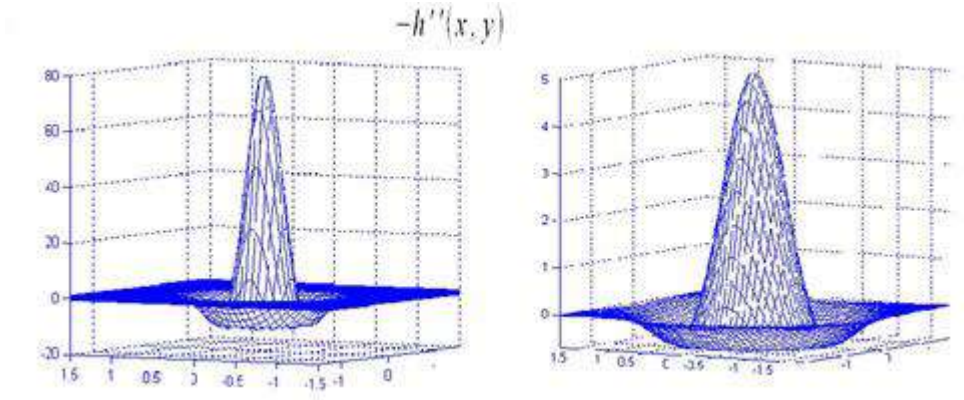


Figure [II.10] : Allure de la fonction Laplacien de Gaussienne pour deux valeurs de σ [1].

Le traitement repose sur cinq paramètres :

- *N* représente la taille du masque (matrice carrée) implantant le filtre LOG. *N* est impair.
- σ permet d'ajuster la taille du chapeau mexicain.
- Δ_x et Δ_y sont les pas d'échantillonnage utilisés pour discrétiser *h''*(*x*,*y*). Généralement $\Delta_x = \Delta_y$
- *S* est le seuil qui permet de sélectionner les contours les plus marqués [1].

Remarque :

Il est à noter que le choix des paramètres *N*, σ et Δ_x ne doit pas se faire de façon indépendante. En effet, le masque, même de taille réduite, doit ressembler à un chapeau mexicain. Le problème ici est le même que celui que l'on rencontre lors de l'échantillonnage d'une fonction gaussienne. Le nombre de points N à considérer doit être tel que l'étendue occupe l'intervalle [-3σ, 3σ].

En fonction du pas d'échantillonnage, l'étendue spatiale vaut : $(N-1)\Delta_x$

Cette étendue peut aussi s'écrire en fonction de σ : $(N-1)\Delta_x = k\sigma$ avec *k* entier.

En prenant par exemple, $\Delta_x = 1$ il s'agit de choisir *N* et σ de sorte que l'étendue du chapeau mexicain soit pertinente. Pour le chapeau mexicain, la valeur de *k* doit être au moins de 4.

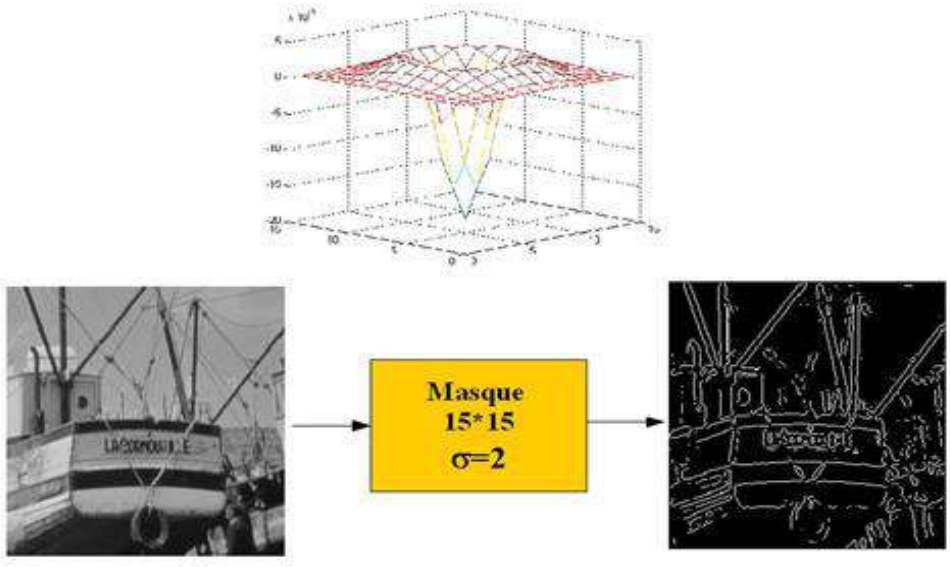


Figure [II.11] : Application du filtre LOG à la détection de contours [1].

Remarque :

Une animation dans l'étude de cas permet d'étudier l'influence du choix de la taille du masque et de l'écart-type σ.

II.4.Conclusion :

Nous avons présenté quelques approches permettant d'extraire les contours d'une image. Ces méthodes sont basées sur la détection des maxima du gradient de l'image ou sur celle des passages par zéro du laplacien. Certaines méthodes sont simples à mettre en œuvre et donnent des résultats assez corrects (filtrage de Sobel). D'autres méthodes comme le Laplacien de Gaussienne ou le filtrage de Canny plus complexes, permettent généralement une meilleure approche des contours. Il existe néanmoins plusieurs autres méthodes non abordées ici qui offrent des perspectives très intéressantes.

Les techniques d'exploration des données

(DATA MINING)

III.1.Introduction :

Les données brutes, malgré leur quantité qui augmente d'une façon exponentielle, n'ont presque aucune valeur, ce qui est le plus important en fait c'est les connaissances pour lesquelles nous sommes tous assoiffés et qui sont obtenus par la compréhension de ces données, mais plus on a de données plus ce processus devient difficile.

De nos jours, les changements de notre environnement sont dénotés par des capteurs qui sont devenus de plus en plus nombreux. Par conséquent, la compréhension de ces données est très importante. *« […] as long as the world keeps producing data of all kinds […] at an ever increasing rate, the demand for data mining will continue to grow* (*Piatestky-Shapiro)»* [12]. D'où la fouille de données devient une nécessité [4].

III.2.Définition du *data mining* :

Data mining traduit généralement en Fouille de données est un ensemble de techniques d'exploration de données permettant d'extraire d'une base de données des connaissances sous la forme de modèles de description afin de décrire le comportement actuel et/ou de prédire le comportement futur du système représenté par ces données.

Data mining se situe à la croisée des bases de données, l'intelligence Artificielle, statistique et l'analyse de données :

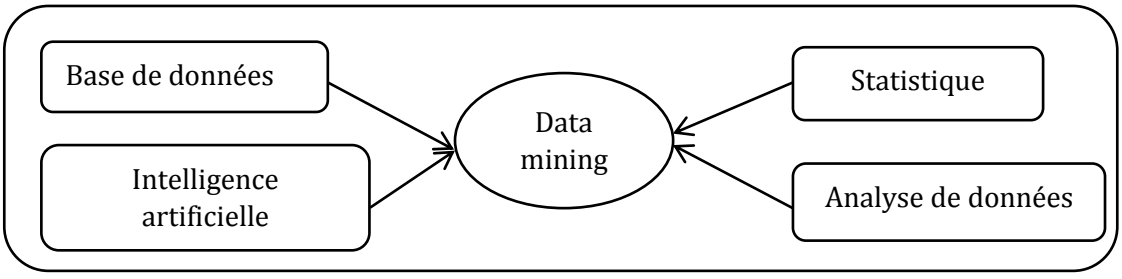


Figure [III.1] : Les Techniques du data mining [3].

Data mining impose d'être capable d'utiliser de manière opérationnelle les résultats des analyses effectuées, souvent dans des délais très courts. Le processus d'analyse doit permettre à l'organisation une réactivité importante.

Les données traitées sont issues de différent systèmes de stockage en place dans l'organisation et sont ainsi hétérogènes, multiples, plus ou moins structurées.

Data mining se propose donc d'extraire des connaissances à partir de grands volumes de données qui peuvent être stockées de manière diverse, dans des bases de données, dans un (ou plusieurs) entrepôt de données (*data warehouse),* mais qui peuvent aussi être récupérées de sources riches plus ou moins structurées comme Internet (*Web*), ou encore en temps réel (retrait d'argent dans un distributeur de billets...).

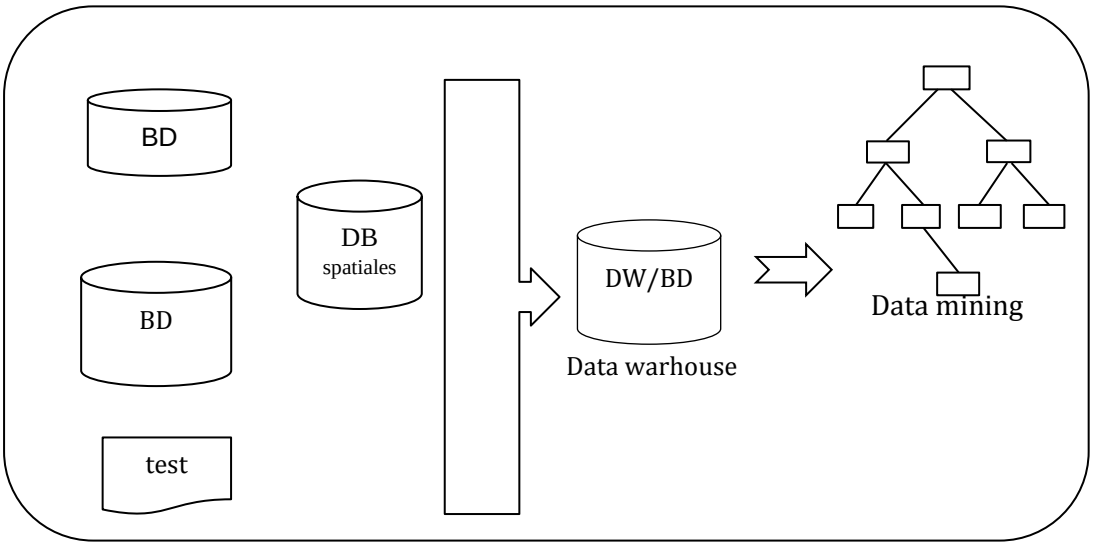


Figure [III.2] : Architecture du *Data Mining* [3].

III.3. Principe et spécificité :

3.1. Stratégie du *Data Mining* :

Les statisticiens utilisent en générale des données expérimentales pour étudier un phénomène, en fixant certaines hypothèses. Par contre, *Data Mining* utilise les données pour trouver toutes les hypothèses valables.

Data mining tente alors de réaliser un arbitrage entre validité scientifique, interprétabilité des résultats et facilité d'utilisation, dans un environnement professionnel où le temps d'étude joue un rôle majeur et où les analystes ne sont pas toujours des statisticiens [3].

3.2. Méthodologie :

Data Mining est un processus d'extraction de connaissances d'un large volume de données comportant les étapes suivantes :

1. Formaliser ou identifier un problème que l'organisation cherche à résoudre en termes de données en cernant les objectifs.

2. Accéder aux données appropriées.

3. Préparer les données en vue des traitements et utilisations futures (collecter, nettoyer, enrichir, coder et normaliser les données).

4. Modéliser les données en leur appliquant des algorithmes d'analyse.

5. Evaluer et valider les connaissances ainsi extraites des analyses.

6. Déployer les analyses pour une utilisation effective.

Ce processus est cyclique, *Data Mining* participe à une meilleure compréhension de l'activité de l'organisation, et à une rationalisation avancée du stockage de l'information et des données. *Data Mining* ne consiste pas en une succession d'études, mais doit bien permettre de capitaliser des connaissances acquises sous forme de connaissance explicites, donc de structurer les contenus nécessaires à l'ingénierie des connaissances [3].

III.4.D*ata mining* sur quels types de données ?

Data mining n'est pas spécifique à un type de médias ou de données. Il est applicable à n'importe quel type d'information. *Data mining* est utilisé et étudié pour les Bases de Données incluant les bases de données relationnelles et les bases de données orientées-objets, *data warehouses*, les bases de données transactionnelles, les supports de données non structurés et semi-structurés comme le *World Wide Web*, les bases de données avancés comme les bases de données spatiales, les bases de données multimédia, les bases de données de séries temporelles et les bases de données textuelles et même fichiers plats [4].

4.1. Les fichiers plats :

Les fichiers plats sont actuellement la source de données la plus commune pour les algorithmes du *Data mining* et particulièrement dans le niveau de recherches. Les fichiers plats sont des fichiers de données simples dans le format texte ou binaire avec une structure connue par l'algorithme du *Data mining* qui va –t– être y appliqué [4].

4.2. Les bases de données relationnelles :

Les algorithmes du *Data mining* appliqués sur des bases de données relationnelles sont plus polyvalents que les algorithmes spécifiquement faits pour les fichiers plats puisqu'ils peuvent profiter de la structure inhérente aux bases de données relationnelles. Le *Data mining* peut profiter du SQL pour la sélection, la transformation et la consolidation, il passe au-delà de ce que le SQL pourrait fournir, comme la prévision, la comparaison, la détection des déviations, etc [4].

4.3. Les data warehouses :

Data Warehouse est un support de données rassemblées de multiples sources de données (souvent hétérogènes) et est destinée à être utilisé dans l'ensemble sous le même schéma unifié. Supposons une entreprise (*OurVideoStore)* qui a des contrats d'exclusivité dans toute l'Amérique du Nord. La plupart des magasins de vidéos appartenant à l'entreprise (*OurVideoStore)* peuvent avoir des bases de données et des structures différentes. Si le directeur de l'entreprise veut accéder aux données de tous les magasins pour prendre des décisions stratégiques, il serait plus approprié si toutes les données étaient stockées dans un seul emplacement avec une structure homogène qui permet l'analyse interactive des données. Autrement dit, les données de différents magasins peuvent être chargées, nettoyées, transformées et intégrées ensemble. Pour faciliter la prise de décisions et les vues multidimensionnelles, *Data Warehouses* sont souvent modelées par une structure de données multidimensionnelle [4].

4.5. Les bases de données multimédia :

Les bases de données multimédia comportent des documents sonores, des vidéos, des images et des médias en textes et audio. Elles peuvent être stockées sur des bases de données orientées objets ou objets relationnelles ou simplement sur un fichier système. Le multimédia est caractérisé par sa haute dimension ce qui rend *data mining* sur ce type de données très

difficile. *Data mining* sur les supports des multimédias requiert exige la vision par ordinateur, l'infographie, l'interprétation des images et les méthodologies de traitement de langages naturels [4].

4.6. Les bases de données spatiales :

Ce sont des bases de données, qu'en plus de leurs données usuelles, elles contiennent des informations géographiques comme les cartes et les positionnements mondiaux ou régionaux. De telles bases de données présentent de nouveaux défis aux algorithmes de *data mining* [4].

4.7. Les bases de données de séries temporelles :

Les bases de données de séries temporelles contiennent des données relatives au temps, comme les données du marché boursier ou les activités enregistrées. Ces bases de données ont couramment un flux continu de nouvelles données entrantes, qui parfois rend l'analyse en temps réel un besoin exigeant. *Data mining* pour ce genre de bases de données est généralement l'étude des tendances et des corrélations entre les évolutions des différentes variables, aussi bien que la prédiction des tendances et des mouvements des variables par rapport au temps [4].

4.8. Le World Wide Web :

Le World Wide Web est le support de données le plus hétérogène et le plus dynamique disponible. Un grand nombre d'auteurs et d'éditeurs contribuent sans arrêt à son accroissement et évolution, et chaque jour un énorme nombre d'utilisateurs accède à ses ressources. Les données dans le World Wide Web sont organisées dans des documents interconnectés.

Ces documents peuvent être des textes, audio, vidéos, données brutes et même des applications. Conceptuellement, le World Wide Web est composé de trois grands composants: le contenu du Web, qui englobe les documents disponibles ; la structure du Web, qui garantit les hyperliens et les relations entre documents ; et l'usage du Web, en décrivant quand et comment les ressources seront accédées. Une quatrième dimension peut être ajoutée concernant la nature dynamique ou l'évolution des documents. *Data mining* pour le World Wide Web, ou le *web mining,* essaie d'aborder toutes ces questions et il est souvent divisé en contenu *Web mining,* la structure *Web mining* et l'usage *Web mining* [4].

III.5.Taxinomie des méthodes :

Afin de construire des modèles à partir des données, *Data Mining* se propose d'utiliser un ensemble d'algorithmes issus de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données). Donc, le but est de trouver des schémas « intéressants » (des patterns ou motifs de conception) selon des critères fixés au départ, et extraire de ces données un maximum de connaissances.

Il existe trois typologies des méthodes qui sont :

5.1. Selon les objectifs :

Selon des objectifs voulus, nous avons quatre types de méthodes :

1. Classification : Consiste à examiner les caractéristiques d'un objet et lui attribuer une classe (attribuer ou non un prêt à un client, établir un diagnostic,…). Donc, la classification consiste à apprendre à classer des objets dans des classes prédéfinies à partir d'exemples déjà classés. L'application de la classification construis un modèle à partir d'un ensemble d'apprentissage, où les éléments sont groupés dans des classes prédéfinies, et ensuite utilise ce modèle pour classifier automatiquement des nouveaux objets dans l'une des classes prédéfinies

2. Prédiction : Consiste à prédire la valeur future d'un attribut en fonction d'autres attributs (prédire la "qualité" d'un client en fonction de son revenu, …). Le but de la prédiction est d'attribuer à des objets une valeur pour une variable continue.

3. Association : Consiste à déterminer les attributs qui sont corrélés. Il s'agit de trouver des similarités ou des associations. L'application principale est "l'analyse du panier de la ménagère ", c'est à dire la recherche d'associations entre produits sur les tickets de caisse. Le but de la méthode est l'étude de ce que les clients achètent pour obtenir des informations sur les clients et pourquoi ils font certains achats. La méthode recherche les produits qui pourront être achetés ensemble. Les domaines applicables sont tous les secteurs pour lesquels il est intéressant de rechercher des groupements potentiels de produits.

4. Segmentation : Consiste à former des groupes homogènes à l'intérieur d'une population. Le cluster ING non supervisée consiste à chercher des groupes (ou classes) d'objets en se basant sur les caractéristiques de ces variables. Ces groupes doivent vérifier deux propriétés.

D'une part, ils ne sont pas prédéfinis, mais découvert automatiquement au cours de l'opération. D'autres part, les clusters / groupes contiennent des objets ayant des caractéristiques similaires et séparent les objets ayant des caractéristiques différentes. On utilise souvent les algorithmes de types nués dynamiques (les centres mobiles, K-mens,. .) ou les cartes auto-organisatrices de *Kohner* [3].

5.2. Selon le type d'apprentissage :

Pour résoudre une problématique avec un processus de *Data Maning*, l'utilisation d'un grand nombre de méthodes et d'algorithmes différents s'impose. On peut distinguer trois grandes familles d'algorithmes :

1. Les méthodes non-supervisées :

L'objectif de la méthode non supervisée est de trouver des relations entre les variables explicatives suffisamment significatives et permettant d'augmenter les connaissances du domaine étudié. Elles permettent de travailler sur un ensemble de données dans lequel aucune des données ou des variables à disposition n'a d'importance particulière par rapport aux autres, c'est-à-dire un ensemble de données dans lequel aucune variable n'est considérée individuellement comme la cible de l'objectif de l'analyse. On les utilise par exemple pour dégager d'un ensemble d'individus des groupes homogènes (typologie), pour construire des normes de comportements et donc des déviations par rapport à ces normes (détection de fraudes nouvelles ou inconnues à la carte bancaire, à l'assurance maladie...), pour réaliser de la compression d'informations (compression d'image).

Parmi les techniques disponibles, nous avons :

- Techniques à base de Réseau de neurones : réseau de Kohner (SOM/TOM) (Carte Auto Adaptative), les réseaux Hebden...
- Techniques utilisées classiquement dans le monde des statistiques : classification ascendante hiérarchique, k-mans et les nuées dynamiques (Recherche des plus proches voisins), les classifications mixtes (Bircher...), les classifications relationnelles...
- Les techniques dites de recherche d'associations :elles sont à l'origine utilisées pour faire de l'analyse dite de panier d'achats ou de séquences, c'est-à-dire pour essayer de savoir parmi un ensemble d'achats effectués par un très grand nombre de clients et de

produits possibles, quels sont les produits qui sont achetés simultanément (pour un supermarché par exemple ; elles sont également appliquées à des problèmes d'analyse de parcours de navigation de site web).: algorithmes apriori, GRI, Carma, méthode ARD...

2. Les méthodes supervisées :

Dans le cadre de l'approche supervisée, les données sont constituées d'un ensemble de caractéristiques, appelées variables exogènes(ou variable explicative), décrivant chaque individu. Chaque individu possède une caractéristique particulière appelée variable endogène (variable à expliquer ou variable cible). L'objectif de la fouille de données supervisée est de trouver des relations entre les variables exogènes permettant d'expliquer et/ou de prévoir le comportement de la variable endogène.

La fouille de données supervisée se décompose en deux étapes : apprentissage d'un modèle sur une partie des données et validation du modèle sur l'autre partie des données.

Les données sont donc divisées en deux ensembles : l'ensemble d'apprentissage et l'ensemble de test. Par exemple, on utilisera ce type de méthode lorsque l'on cherchera à comprendre pourquoi un individu a acheté un produit plutôt qu'un autre, pourquoi un individu a répondu favorablement à une opération de marketing direct, pourquoi un individu a contracté une maladie particulière.

Parmi les techniques disponibles, nous avons :

- Techniques à base d'arbres de décision: CART, CHAID, ECHAID, QUEST, C5, C4.5, les forêts d'arbres...
- Techniques statistiques de régressions linéaires et non linéaires au sens large : Régression linéaire, Régression linéaire multiple, Régression logistique binaire ou multinomiale, Probité binaire, multinomial ou ordonné, Tobit, Gauchit, modèle Gamma, binomial négatif, log-log, Analyse discriminante linéaire ou quadratique, régression de coxa, modèle linéaire généralisé, régression PLS, régressions non paramétrique, équations structurelles.
- Techniques à base de Réseau de neurones : Perceptron mono ou multicouches avec ou sans rétro propagation des erreurs, les réseaux à fonction radiale de base...
- Techniques à base d'algorithme génétique.

- Techniques le raisonnement par cas.
- Le Filtrage collaboratif à base d'inférence bayésienne (Réseau bayésien).

3. Les méthodes de réduction de données :

Elles permettent de réduire un ensemble de données volumineux à un ensemble de taille plus réduite, épuré de ce que l'on considérera comme de l'information non pertinente ou non signifiante, comme du bruit. Elles sont ainsi très souvent, mais pas systématiquement, utilisées en amont des techniques supervisées ou non supervisées. Elles sont notamment très complémentaires des techniques non supervisées classiquement utilisées dans le domaine de statistiques.

- Techniques d'analyse factorielle : Analyse en composantes principales, analyse factorielle des correspondances, analyse des correspondances multiples, analyses factorielles (maximum de vraisemblance, moindres carrés non pondérés, avec ou sans rotation orthogonale ou oblique)...
- Techniques de positionnement [3] : positionnement multidimensionnel....

5.3- Selon les types de modèles obtenus :

1. Modèles prédictifs : Utilisent les données avec des résultats connus pour développer des modèles permettant de prédire les valeurs d'autres données (modèle permettant de prédire les clients qui ne rembourseront pas leur crédit, classification, prédiction). Ils s'intéressent à la recherche de structures ou modèles destinés à la prévision de certaines quantités.

2. Modèles descriptifs : Proposent des descriptions des données pour aider à la prise de décision. Les modèles descriptifs aident à la construction de modèles prédictifs (segmentation, Association). Ils s'intéressent aux individus ainsi qu'aux groupes homogènes qu'ils peuvent former [3].

III.6.Les étapes du processus de *data mining* :

1- Collecte des données : la combinaison de plusieurs sources de données, souvent hétérogènes, dans une base de données.

2- Nettoyage des données : la normalisation des données : l'élimination du bruit (les attributs ayant des valeurs invalides et les attributs sans valeurs).

3- Sélection des données : Sélectionner de la base de données les attributs utiles pour une tâche particulière du *data mining*.

4- Transformation des données : le processus de transformation des structures des attributs pour être adéquates à la procédure d'extraction des informations.

5- Extraction des informations (*Data mining):* l'application de quelques algorithmes du *Data Mining* sur les données produites par l'étape précédente (*Knowledge Discovery in Databases*, ou KDD).

6- Visualisation des données : l'utilisation des techniques de visualisation (histogramme, camembert, arbre, visualisation 3D) pour exploration interactive de données (la découverte des modèles de données).

7- Evaluation des modèles : l'identification des modèles strictement intéressants en se basant sur des mesures données [4].

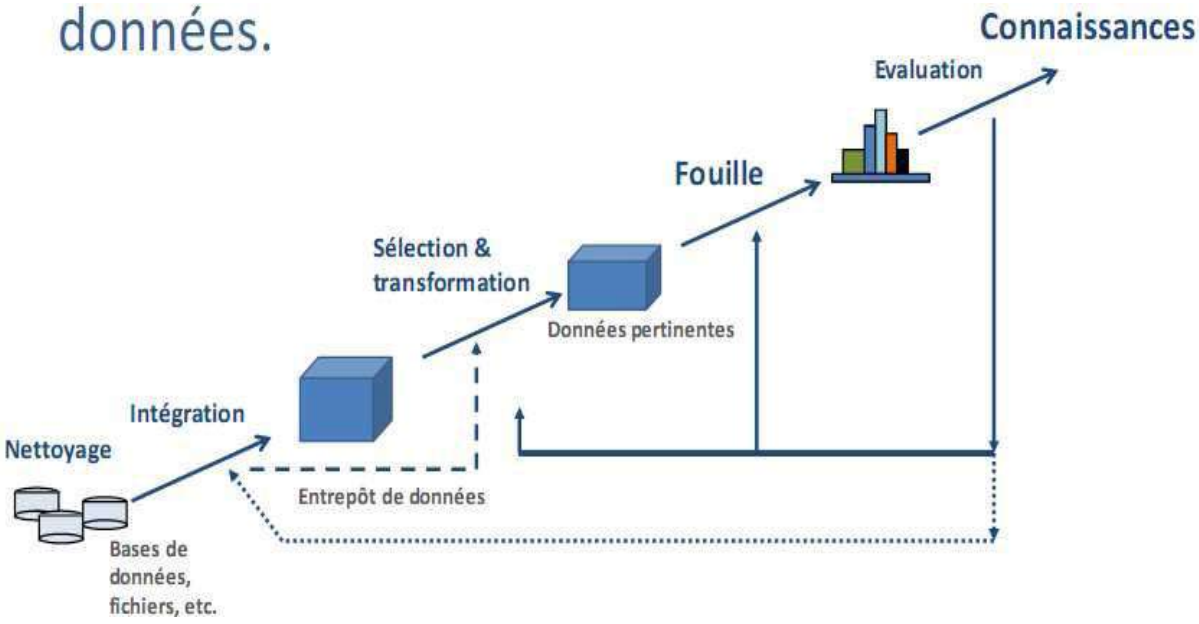


Figure [III.3]: Le processus de *data mining* [4].

III.7.Techniques du *data mining* :

Pour effectuer les taches du *Data mining* il existe plusieurs techniques issus de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données) afin de faire

apparaître des corrélations cachées dans des gisements de données pour construire des modèles à partir de ces données [4].

7.1. Les réseaux de neurones :

Un réseau de neurones est un modèle de calcul dont le fonctionnement vise à simuler le fonctionnement des neurones biologiques, il est constitué d'un grand nombre d'unités (neurones) ayant chacune une petite mémoire locale et interconnectées par des canaux de communication qui transportent des données numériques. Ces unités peuvent uniquement agir sur leurs données locales et sur les entrées qu'elles reçoivent par leurs connections. Les réseaux de neurones sont capables de prédire de nouvelles observations (sur des variables spécifiques) à partir d'autres observations (soit les même ou d'autres variables) après avoir exécuté un processus d'apprentissage sur des données existantes.

La phase d'apprentissage d'un réseau de neurones est un processus itératif permettant de régler les poids du réseau pour optimiser la prédiction des échantillons de données sur lesquelles l'apprentissage été fait. Après la phase d'apprentissage le réseau de neurones devient capable de généraliser [4].

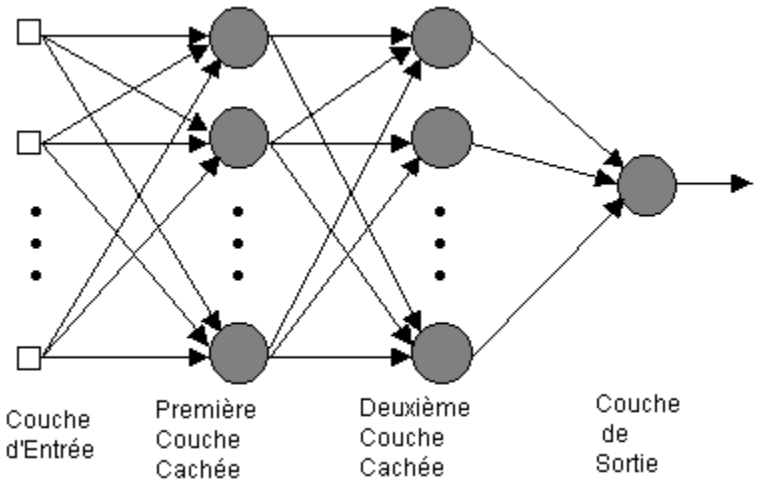


Figure [III.4] : Réseau de neurones [4].

7.1.1- Avantages et inconvénients :

- **Avantages**

Les réseaux de neurones sont théoriquement capables d'approximer n'importe quelle fonction continue et ainsi le chercheur n'a pas besoin d'avoir aucunes hypothèses du modèle sous-jacent.

- **Inconvénients**

Généralement les réseaux de neurones ne sont pas souvent utilisées dans les tâches du *data mining* parce qu'ils produisent des modèles souvent incompréhensibles et demande un long temps d'apprentissage.

7.2. Les arbres de décision :

Les arbres de décisions sont des outils d'aide à la décision qui permettent selon des variables discriminantes de répartir une population d'individus en groupes homogènes en fonction d'un objectif connu. Les arbres de décision sont des outils puissants et populaires pour la classification et la prédiction. Un arbre de décision permet à partir des données connues sur le problème de donner des prédictions par réduction, niveau par niveau, du domaine des solutions.

Chaque nœud interne d'un arbre de décision permet de répartir les éléments à classifier de façon homogène entre ses différents fils en portant sur une variable discriminante de ces éléments. Les branches qui représentent les liaisons entre un nœud et ses fils sont les valeurs discriminantes de la variable du nœud. Et en fin, les feuilles d'un arbre de décision représentent les résultats de la prédiction des données à classifier [4].

7.2.1 -Avantages et inconvénients

- **Avantages**
  - ❖ Les arbres de décision sont capables de produire des règles compréhensibles.
  - ❖ Les arbres de décision effectuent la classification sans exiger beaucoup de calcul
  - ❖ Les arbres de décision sont en mesure de manipuler à la fois les variables continues et catégorielles.

- **Inconvénients**
  - ❖ Manque de performance dans le cas de plusieurs classes; les arbres deviennent très complexes et ne sont pas nécessairement optimaux.
  - ❖ Demande beaucoup de temps de calcul lors de la construction (le choix du meilleur partitionnement) et l'élagage (la comparaison de sous-arbres).
  - ❖ Moins bonnes performances concernant les prédictions portant sur des valeurs numériques.

7.3. Les algorithmes génétiques :

Un algorithme génétique se constitue d'une catégorie de programmes dont le principe est la reproduction des mécanismes de la sélection naturelle pour résoudre un problème donné. L'optimisation des problèmes combinatoires et surtout les problèmes dits NP-complets (dont le temps de calcul croit de façon non polynomiale avec la complexité du problème) est l'objectif principale des algorithmes génétiques, ils sont particulièrement adaptés à ce type de problèmes. Ces algorithmes constituent parfois une alternative intéressante aux réseaux de neurones mais sont le plus souvent complémentaires [4].

7.3.1- Principe de base des algorithmes génétiques :

Le principe de fonctionnement d'un algorithme génétique est le suivant :

1. Codage du problème sous forme d'une chaîne binaire.

2. Génération aléatoire d'une population. Celle-ci contient un pool génétique qui représente un ensemble de solutions possibles.

3. Calcul d'une valeur d'adaptation pour chaque individu. Elle sera fonction directe de la proximité des différents individus avec l'objectif, on parle ici d'évaluation (fitness).

4. Sélection des individus doit se reproduire en fonction de leurs parts respectives dans l'adaptation globale.

5. Croisement des génomes des parents.

6. Sur la base de ce nouveau pool génétique, on repart à partir du point ***3***.

7.3.2-Codage d'un algorithme génétique :

Avant de pouvoir utiliser un algorithme génétique pour résoudre un problème, il faut trouver un moyen pour encoder une solution potentielle à ce problème (les chromosomes), le codage consiste alors à choisir les structures de données qui coderont les gènes. Il existe différentes manières de le faire :

- **Codage binaire :** ce type de codage se base sur le principe de coder la solution selon une chaîne de bits (0 ou 1).
- **Codage à caractères multiples :** les chromosomes d'un algorithme génétique peuvent être codés d'une autre manière qui est le codage à l'aide de caractères. Souvent, ce type de codage est plus naturel que son précédent.
- **Codage sous forme d'arbre :** il utilise une structure arborescente avec une racine (parent) de laquelle peuvent être issus un ou plusieurs fils (descendants).

7.3.3-Avantages et inconvénients :

- **Avantages**
  - Ils utilisent l'évaluation de la fonction objective sans prendre en compte sa nature ce qui lui donne plus de souplesse et un large domaine d'application.
  - Ils sont dotés de parallélisme car ils travaillent sur plusieurs points en même temps il s'agit des individus de la population.
  - L'utilisation de règles de transition probabilistes de croisement et de mutation permet dans certains cas d'éviter des optimums locaux et d'aller vers un optimum global.

- **Inconvénients**
  - Temps de calcul très élever car ils nécessitent de nombreux calculs particulièrement au niveau de la fonction d'évaluation.
  - Difficiles à mettre en œuvre à cause

Des paramètres parfois difficiles à déterminer comme la taille de la population ou le taux de mutation. Ce qui implique la nécessité de plusieurs essais car le succès de l'évolution en dépend, ce qui limite encore l'efficacité de l'algorithme.

Du choix de la fonction d'évaluation qui est critique, elle doit prendre en compte les bons paramètres du problème. Elle doit donc être choisie avec soin.

- ❖ Il est impossible d'être sure que la solution obtenue après un nombre fini de générations soit la meilleure, on peut seulement être sûr que l'on s'est approché de la solution optimale.
- ❖ Problème des optimums locaux : lorsqu'une population évolue, il se peut que certains individus deviennent majoritaires. À ce moment, il se peut que la population converge vers cet individu et s'écarte ainsi d'individus plus intéressants mais trop éloignés de l'individu vers lequel la population converge [5].

7.4. Les règles associatives :

Les règles associatives sont des règles extraites d'une base de données transactionnelles et qui décrivent des associations entre certains éléments. Elles sont fréquemment utilisées dans le secteur de la distribution des produits où la principale application est l'analyse du panier de la ménagère (*Market Basket Analysis*) dont le principe est l'extraction d'associations entre produits sur les tickets de caisse. Le but de la méthode est l'étude de ce que les clients achètent pour obtenir des informations sur qui sont les clients et pourquoi ils font certains achats. La méthode recherche quels produits tendent à être achetés ensemble. La méthode peut être appliquée à tout secteur d'activité pour lequel il est intéressant de rechercher des groupements potentiels de produits ou de services : services bancaires, services de télécommunications, par exemple. Elle peut être également utilisée dans le secteur médical pour la recherche de complications dues à des associations de médicaments ou à la recherche de fraudes en recherchant des associations inhabituelles.

Une règle d'association est de la forme : Si condition alors résultat. Dans la pratique, nous nous limitons généralement à des règles où la condition se présente sous la forme d'une conjonction d'apparition d'articles et le résultat se constitue d'un seul article. Par exemple, une règle à trois articles sera de la forme : Si X et Y alors Z ; règle dont la sémantique peut-être énoncée : Si les articles X et Y apparaissent simultanément dans un achat alors l'article Z apparaît [4].

7.4.1- Avantages et inconvénients

- **Avantages**
  - ❖ Résultats clairs : règles faciles à interpréter.
  - ❖ Simplicité de la méthode et des calculs (calculs élémentaires des fréquences d'apparition).

- ❖ Aucune hypothèse préalable (Apprentissage non supervisé).
- ❖ Méthode facile à adopter aux séries temporelles.

▪ **Inconvénients**

- ❖ La méthode est coûteuse en temps de calcul.
- ❖ Qualité des règles : production d'un nombre important de règles triviales (des règles évidentes qui, par conséquent, n'apportent pas d'information) ou inutiles (des règles difficiles à interpréter provenant de particularités propres à la liste des achats ayant servie à l'apprentissage).
- ❖ Méthode non efficace pour les articles rares.

7.5. L'algorithme des k-Plus proches voisins :

L'algorithme des k plus proches voisins (K-PPV, k *nearest neighbor* en anglais ou *kNN*) est un algorithme de raisonnement à partir de cas qui est dédié à la classification qui peut être étendu à des tâches d'estimation. Le but de cet algorithme est de prendre des décisions en se basant sur un ou plusieurs cas similaires déjà résolus en mémoire. Dans ce cadre, et Contrairement aux autres méthodes de classification (arbres de décision, réseaux de neurones, algorithmes génétiques, …etc.) l'algorithme de KNN ne construit pas de modèle à partir d'un échantillon d'apprentissage, mais c'est l'échantillon d'apprentissage, la fonction de distance et la fonction de choix de la classe en fonction des classes des voisins les plus proches, qui constituent le modèle [4] [5].

7.5.1- Algorithme de classification par k-PPV

Paramètre : le nombre k de voisins

Donnée : un échantillon de m exemples et leurs classes

La classe d'un exemple X est c(X)

Entrée : un enregistrement Y

1. Déterminer les k plus proches exemples de Y en calculant les distances

2. Combiner les classes de ces k exemples en une classe c

Sortie : la classe de Y est c(Y)=c.

7.5.3- Avantages et inconvénients

- **Avantages**
  - La qualité de la méthode s'améliore en introduisant de nouvelles données sans nécessiter la reconstruction d'un modèle. Ce qui représente une différence majeure avec des méthodes telles que les arbres de décision et les réseaux de neurones.
  - La clarté des résultats : la classe attribuée à un objet peut être expliquée en exhibant les plus proches voisins qui ont amené à ce choix.
  - La méthode peut s'appliquer à tout type de données même les données complexes tels que des informations géographiques, des textes, des images, du son. C'est parfois un critère de choix de la méthode PPV car les autres méthodes traitent difficilement les données complexes. Nous pouvons noter, également, que la méthode est robuste au bruit.
  - Facile à mettre en œuvre.
- **Inconvénients**
  - Temps de classification : la méthode ne nécessite pas d'apprentissage ce qui implique que tous les calculs sont effectués lors de la classification. Contrairement aux autres méthodes qui nécessite un apprentissage (éventuellement long) mais qui sont rapides en classification.
  - Méthode donnera de mauvais résultats Si le nombre d'attributs pertinents est faible relativement au nombre total d'attributs, car la proximité sur les attributs pertinents sera noyée par les distances sur les attributs non pertinents.
  - Les performances de la méthode dépendent du choix de la distance, du nombre de voisins et du mode de combinaison des réponses des voisins [4].

7.6. L'algorithme des k-moyennes (K-Means) :

L'algorithme des K-moyennes est dédié aux tâche de *clustering*, il permet de diviser une population donnée en K groupes homogènes appelés clusters. Le nombre de clusters K est déterminé par l'utilisateur selon ses attentes [4].

7.6.1- Principe de fonctionnement :

Après avoir déterminé un nombre K de clusters nous positionnons les K premiers points (appelés graines) au hasard (nous utilisons en général les K premiers enregistrements).

Chaque enregistrement est affecté à la graine dont il est plus proche (en utilisant la fonction de distance). A la fin de la première affectation, la valeur moyenne de chaque cluster est calculée et la graine prend cette nouvelle valeur. Le processus est répété jusqu'à stabilisation des clusters [7][4].

7.7. Algorithme de clustering par K-Means:

L'algorithme *k-means* est en 4 étapes :

1. Choisir k objets formant ainsi k clusters
2. Affecter chaque objet O au cluster Ci de centre Mi tel que *dist* (O,Mi) est minimal
3. Recalculer Mi de chaque cluster (le barycentre)
4. Aller à l'étape 2 s'il faut faire une affectation

7.7.1- Evaluation

Le but de la technique des k-*means* est le regroupement par similitude des populations statiques, ce qui nécessite de déterminer la qualité des différents clusters. Une évaluation de la qualité pourrait *co nsister* à l'étude de la variance de cette population. Alors, nous pouvons dire qu'un cluster possédant d'une part population et d'autre part une variance faible est un cluster solide.

Il est nécessaire d'effectuer d'autres évaluations dans les cas suivants :

- ❖ Si la population d'un cluster est trop faible, il sera préférable de grouper ce cluster avec un autre.
- ❖ Si un cluster est trop dominant, il pourrait être valable de diviser la population en deux (dans et hors cluster) et de relancer le processus pour chaque sous-groupe **[7]**.

7.7.2- Avantages et inconvénients

- **Avantages**
  - ❖ La méthode résolve une tâche non supervisée, donc elle ne nécessite aucune information sur les données.
  - ❖ Technique facile à mettre en œuvre.

- ❖ La méthode est applicable à tout type de données (mêmes textuelles), en choisissant une bonne notion de distance.

- **Inconvénients**
  - ❖ La difficulté de trouver une bonne fonction de distance.
  - ❖ Un bon choix du nombre k est nécessaire, un mauvais choix de k produit de mauvais résultats.
  - ❖ La difficulté d'expliquer certains clusters (i.e. attribuer une signification aux groupes constitués) [7] [4].

Les techniques de fouille de données présentées ci-dessus représentent une partie des techniques existantes pour l'extraction des données, et ce grand nombre de techniques est dû qu'ils n'ont pas tous le même objet, et qu'aucun n'est optimal dans tous les cas et qu'ils s'avèrent en pratique complémentaires les uns des autres et qu'en les combinant intelligemment (en construisant ce que l'on appelle des modèles de modèles ou méta modèles) il est possible d'obtenir des gains de performance très significatifs.

III.8.Catégorisation des systèmes du *data mining* :

Les systèmes de *data mining* peuvent être catégorisés selon plusieurs critères. Parmi les catégorisations existantes nous citons :

8.1. Classification selon le type de données à explorés : dans cette classification les systèmes de *data mining* sont regroupés selon le type des données qu'ils manipulent tel que les données spatiales, les données de séries temporelles, les données textuelles et le Word Wide Web, etc [4].

8.2. Classification selon les modèles de données avancés : cette classification catégorise les systèmes de *data mining* en se basant sur les modèles de données avancés tel quelles bases de données relationnelles, les bases de données orientées objets, les *data warehouses*, les bases de données transactionnelle, etc [4].

8.3. Classification selon le type de connaissance à découvrir : cette classification catégorise les systèmes de *data mining* en s'appuyant sur le type de connaissance à découvrir ou les tâches de *data mining* tel que la classification, l'estimation, la prédiction, etc.

8.4. Classification selon les techniques d'exploration utilisées : cette classification catégorise les systèmes de *data mining* suivant l'approche d'analyse de données utilisés la

reconnaissance des formes, les réseaux neurones, les algorithmes génétiques, les statistiques, la visualisation, orienté-base de données ou orienté-*data warehouse*, etc [4].

III.9.Domaines d'application du *data mining* :

La technologie de *data mining* a une grande importance économique grâce aux possibilités qu'elle offre pour optimiser la gestion des ressources (humaines et matérielles). Les domaines d'application actuels du *data mining* sont les suivants :

9.1. Le data mining dans le secteur bancaire :

Le secteur bancaire est à la tête de tous les autres domaines industriels grâce son utilisation des techniques du *Data mining* dans ses grandes bases de données clients. Bien que les banques ont employées des outils d'analyse statistiques avec un peu de succès pendant plusieurs années, les modèles précédemment invisibles des comportements des clients deviennent maintenant plus clair à l'aide des nouveaux outils du *data mining.*

Quelques applications du *data mining* dans ce domaine sont :

- Prédire la réaction des clients aux changements des taux d'intérêt.
- Identifier les clients qui seront les plus réceptifs aux nouvelles offres de produits.
- Identifier les clients "fidèles".
- Déterminer les clients qui posent le risque le plus élevé de manquer à leurs engagements aux prêts.
- Détecter les activités frauduleuses dans les transactions par cartes de crédit.
- Prédire les clients qui sont susceptibles de changer leurs cartes d'affiliation au cours du prochain trimestre.
- Déterminez les préférences des clients pour les différents modes de transaction à savoir par le biais de guichets ou par l'intermédiaire de cartes de crédit, etc.

9.2. Le data mining dans la bio-informatique et la biotechnologie

La Bio-informatique est un domaine de recherche en développement rapide, qui a des racines aussi bien dans la biologie que dans la technologie d'informations. Quelques applications du *data mining* dans ce domaine sont :

- La prédiction les structures de différentes protéines.
- La détermination de la complexité des structures de plusieurs médicaments.

9.3. Le data mining dans le marketing direct et le collecte de fonds

Le marketing direct est un ensemble de techniques permettant d'identifier les consommateurs (particuliers et entreprises) d'un produit stockés dans une base de données, de leur adresser directement et individuellement une proposition commerciale, afin d'obtenir une réponse directe, à laquelle l'entreprise répondra tout aussi directement.

Le Marketing Direct est le premier domaine qui a employé les outils du *data mining* pour son profit.

Les différentes façons dont lesquelles le *data mining* peut aider dans le marketing direct sont :

- Les règles d'association (Analyse du panier de la ménagère) permettent de décider l'emplacement et la promotion des produits dans un magasin et de comparer les résultats entre les différents magasins, entre les clients dans les différents groupes démographiques, entre les différents jours de la semaine, les différentes saisons de l'année, etc. pour augmenter ses ventes.
- Les règles d'association prédictives peuvent être utilisées pour identifier les séries des produits qui s'achètent ensembles. *data mining* peut être utilisé dans la collecte de fonds (par exemple : les organisations bénévoles, collecte de fonds dans les élections).
- Rassembler des données illimitées sur les différents donateurs, les bénévoles, les perspectives, les agences et les membres des différentes fiducies dans une base de données centrale. Ces données sont gérées et utilisées par des outils du *data mining* pour construire et améliorer les relations à long terme avec eux.
- Définition des classes des récompenses, les critères d'attribution et permettre au système de se référer aux données pour générer automatiquement des prix comme une reconnaissance des efforts exceptionnels des donateurs.

9.4. Le data mining dans la détection de fraude

La fouille de données est largement appliquée dans des processus de détection de fraude divers tel que :

- Détection de fraude de cartes de crédits.
- Détection de fraude dans les listes des électeurs en utilisant les réseaux de neurones en combinaison avec *data mining.*

- La détection des fraudes dans les demandes de passeport par la conception d'un système de diagnostic par apprentissage en ligne.
- Détection de fausses demandes de remboursement médicale [4].

9.5. Le data mining dans la gestion de données scientifiques

Quelques exemples de la fouille de données dans l'environnement scientifique sont :

9.5.1-Les études sur les changements climatiques du globe: Il s'agit d'un domaine de recherche *chaud* et est essentiellement un exercice de vérification axée sur l'exploitation. À travers les données climatiques qui ont été recueillies au fil des siècles et qui sont en train d'être étendues dans le passé lointain et, en même temps, à travers des activités telles que l'analyse des échantillons de carottes de glace de l'Antarctique, des différents modèles de prédiction ont été proposées pour les futures conditions climatiques.

9.5.2- Les études sur les bases de données de géophysique à l'Université d'Helsinki:

Ils ont publié une analyse scientifique des données sur l'agriculture et l'environnement. En conséquence, ils ont optimisé le rendement des cultures, tout en réduisant au minimum les ressources fournies. Afin de réduire au minimum les ressources, ils ont identifié les facteurs qui influent sur le rendement des cultures, comme les engrais chimiques et les additifs (phosphate), Le contenu d'humidité et le type du sol.

9.6. Le data mining dans le secteur des assurances

Les compagnies d'assurance peuvent bénéficier des méthodes du *data mining*, qui aident les entreprises à réduire les coûts, augmenter les profits, de conserver les clients actuels, d'acquérir de nouveaux clients, et développer de nouveaux produits. Cela peut être fait par le biais de:

- Evaluation du risque d'un bien assuré prenant en compte les caractéristiques du bien et de son propriétaire.
- Formulation des modèles statistiques des risques d'assurance.
- Utilisation du modèle de l'exploitation de Poisson / Log-normale afin d'optimiser les polices d'assurance [4].

9.7. Le data mining dans la télécommunication

À nos jours, toute activité de télécommunication a utilisé une la technique de *data mining*.

- Analyse des achats de services de télécommunications.
- Prédiction de modèles d'appels téléphoniques.
- Gestion des ressources et de trafic réseau.
- Automatisation de la gestion du réseau et de la maintenance en utilisant l'intelligence artificielle pour diagnostiquer et réparer les problèmes de transmission du réseau, etc.

9.8. Le data mining dans la médecine et la pharmacie

Quelques exemples de l'usage médicaux et pharmaceutiques des techniques de *Data Mining* pour l'analyse de bases de données médicales.

- Prédiction de présence de maladies et/ou de complications.
- Le choix d'un traitement pour le cancer.
- Choix des antibiotiques pour des infections.
- Le choix d'une technique particulière (de sutures, matériel de suture, etc.) dans une des procédures chirurgicales.
- Approvisionnement des médicaments les plus fréquemment prescrits.

9.9. Le data mining dans le commerce au détail

Les techniques du *data mining* ont été très utiles pour le CRM (Customer Relationship Marketing) en développant des modèles pour :

- La prédiction de la propension du client à acheter.
- L'évaluation des risques pour chaque transaction.
- Connaitre la distribution et l'emplacement géographique des clients.
- L'analyse de la fidélité des clients dans les opérations à base de crédit.
- Évaluation de la menace concurrentielle dans une région.

9.10. Le data mining dans le e-commerce et le World Wide Web

Quelques façons d'utilisation des outils du *data mining* dans le e-commerce sont :

- En formulant des tactiques du marché dans les opérations de business.

- En automatisant des interactions d'affaires avec des clients, pour que les clients puissent traiter avec tous les acteurs dans la chaîne d'approvisionnement.

La détermination de la taille de *world Wide Web* est extrêmement difficile. En 1999, elle a été estimée à 350 milliards de pages avec un taux de croissance de 1 million de pages / jour. En considérant le World Wide Web comme la plus grande collection de bases de données, le *Web mining* peut être fait par les façons susdites.

9.11. Le data mining dans le marché boursier et l'investissement

L'évolution rapide de la technologie informatique au cours des dernières décennies a facilité l'investissement par des professionnels (et des amateurs), avec la possibilité d'accéder et d'analyser d'énormes quantités de données financières. Les outils du *data mining* sont utilisés pour :

- Aider les spécialistes du marché boursier à prédire les mouvements du prix des actions.
- *Data mining* des anciens des prix et des variables liées aide à découvrir des anomalies de marché boursier comme le scandale *hawala.*

9.12. Le data mining dans l'analyse de chaîne d'approvisionnement

Les techniques du *data mining* ont trouvé une large application dans l'analyse des chaînes d'approvisionnements. Des exemples d'utilisation du *data mining* par les fournisseurs sont :

- Analyser le processus des données pour gérer l'évaluation d'acheteur.
- L'extraction des données de paiement avec l'utilité de mettre à jour la politique tarifaire.

III.10.Conclusion :

Data mining est l'extraction d'informations prédictives cachés dans de grandes base de données. C'est une technologie nouvelle et puissante qui donne la possibilité aux de se concentrer sur les informations les plus importantes dans leurs *data warehouses*. Les outils du *data mining* peuvent prédire les futurs tendances et actions, permettant de prendre les bonnes décisions. C'est ce qui rend *data mining* la technologie la plus importante.

Evaluation des Deux Algorithmes ACP et ADL.

IV.1.Introduction :

Dans ce chapitre on va voir des applications pour la détection des contours et des quelques opérations de gradient, ainsi on fait une évaluation des deux techniques PCA et LDA pour la reconnaissance de visage, utilisant les bases de données ORL et IFD. Ces application est fait avec le langage Matlab.

IV.2.Outils de Développements :

IV.2.1. Matlab 7.6.0 (R2008a) :

Matlab et son environnement interactif est un langage de haute niveau qui permet l'exécution de tâches nécessitant une grande puissance de calcul et dont la mise en œuvre sera bien plus simple et rapide qu'avec des langages de programmation traditionnels tel que le C, C++, il dispose de plusieurs boites à outils en particuliers celle de traitement d'images qui propose un ensemble d'algorithme et d'outils graphiques de référence pour le traitement, l'analyse, la visualisation et le développement d'algorithme de traitement d'images.

IV.3.Filtrage et Détections Des Contours EVALUATION :

A. Lecture et Affichage d'image Numérique:

```
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                       read and display image                           %
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
im = imread('lena.bmp');
figure(1); imshow(im); hold on;
[x1, y1]=ginput(2); plot(x1, y1, '+r');
[x2, y2]=ginput(2); plot(x2, y2, '+r');
xOffs=x1; yOffs=y1; w=x2-x1+1; h=y2-y1+1;

% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                        access image data
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
im2 = double(im(xOffs:xOffs+w-1, yOffs:yOffs+h-1));
figure(2); colormap('gray');
subplot(2,2,1); stem(im2(1,:), '.'); title('row 1');
subplot(2,2,2); stem(im2(:,1), '.'); title('col 1');
subplot(2,2,3); imagesc(im2);
```

```
subplot(2,2,4); meshz(im2);

% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                        point operations- histogram equalisation
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
h1 = imhist(im);
im3 = histeq(im);
h3=imhist(im3);

figure(3);
subplot(2,2,1); imhist(im);
subplot(2,2,2); imshow(im);
subplot(2,2,3); imhist(im3);
subplot(2,2,4); imshow(im3);
```

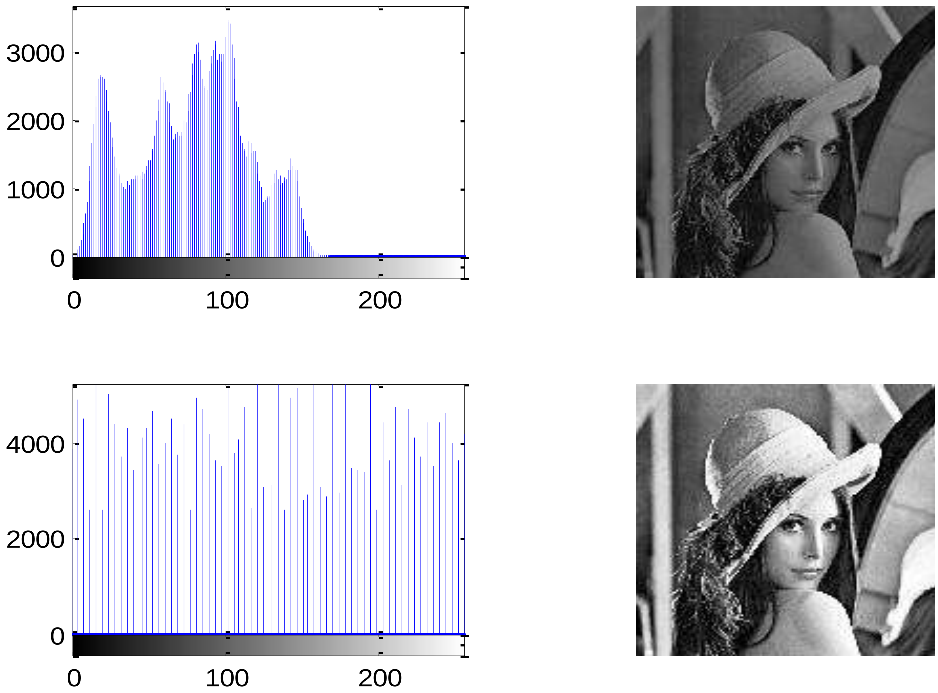


Figure [IV.1]: Affichage et histogramme d'image.

B. Conversion des Couleurs:

```
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                            color conversion
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
input('\n color conversion...');
im1 = imread('para_island.bmp');
im2 = imread('Eagle.bmp');
subplot(2,2,1); imshow(im1);
subplot(2,2,2); imshow(rgb2gray(im1));
subplot(2,2,3); imshow(im2);
subplot(2,2,4); imshow(rgb2gray(im2));
```



Figure [IV.2]: Conversion des couleurs.

C. Operations des Filtrages :

```
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                     filter operations - smoothing
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% build a Gaussian smooth filter:
% gf1 = computeGaussianKernel(2);
% gf2 = computeGaussianKernel(8);
% gf3 = computeGaussianKernel(16);

gf1 = gausswin(4);
gf2 = gausswin(8);
gf3 = gausswin(16);


im_gf1 = imfilter(im, gf1, 'symmetric', 'conv');
im_gf2 = imfilter(im, gf2, 'symmetric', 'conv');
im_gf3 = imfilter(im, gf3, 'symmetric', 'conv');
figure(4);
subplot(2,2,1), imshow(im); title('original lena');
subplot(2,2,2), imshow(im_gf1); title('blured, gaussian, std 4, ');
subplot(2,2,3), imshow(im_gf2); title('blured, gaussian, std 8, ');
subplot(2,2,4), imshow(im_gf3); title('blured, gaussian, std 16,');


% laplacian filter - edge detection:
kLap1 = fspecial('laplacian', 0.2);
kLap2 = fspecial('laplacian', 0.5);
kLap3 = fspecial('laplacian', 0.8);
```

```
im_lap1 = imfilter(im, kLap1, 'symmetric', 'conv');
im_lap2 = imfilter(im, kLap2, 'symmetric', 'conv');
im_lap3 = imfilter(im, kLap3, 'symmetric', 'conv');

figure(5);
subplot(2,2,1), imshow(im); title('original lena');
subplot(2,2,2), imshow(im_lap1); title('alpha=0.2');
subplot(2,2,3), imshow(im_lap1); title('alpha=0.5');
subplot(2,2,4), imshow(im_lap1); title('alpha=0.8');


% edge detector
h1 = [-1 -1 -1; -1 8 -1; -1 -1 -1];
h2 = [-1 -1 -1; 2 2 2; -1 -1 -1];
h3 = [-1 2 -1; -1 2 -1; -1 2 -1];
h4 = [-1 -1 2; -1 2 -1; 2 -1 -1];
h5= h4';
figure(6);
subplot(2,3,1); imshow(im); title('original lena');
subplot(2,3,2); im1 = imfilter(im, h1); imshow(im1); title('h1');
subplot(2,3,3); im2 = imfilter(im, h2); imshow(im2); title('h2');
subplot(2,3,4); im3 = imfilter(im, h3); imshow(im3); title('h3');
subplot(2,3,5); im4 = imfilter(im, h4); imshow(im4); title('h4');
subplot(2,3,6); im5 = imfilter(im, h5); imshow(im5); title('h5');
```



Figure [IV.3]: Opération de Filtrage Gaussien.

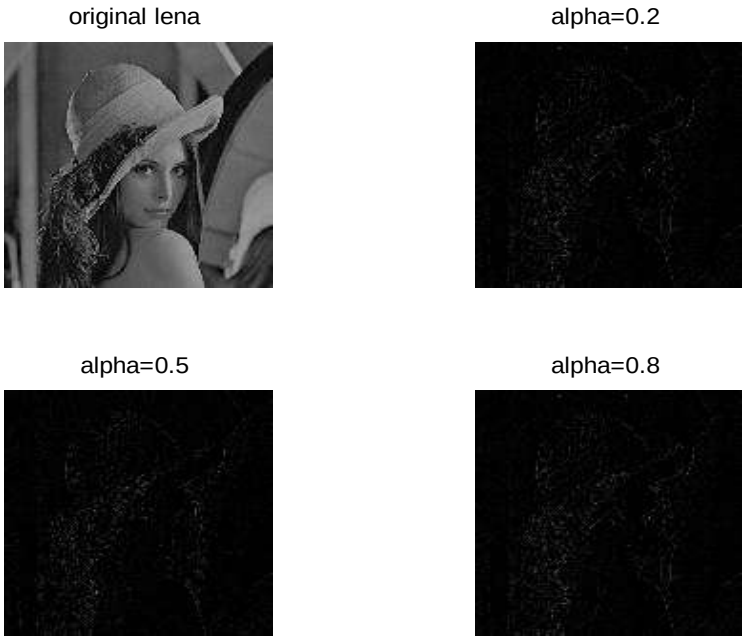


Figure [IV.4]: Filtrage pour les déférents alphas.

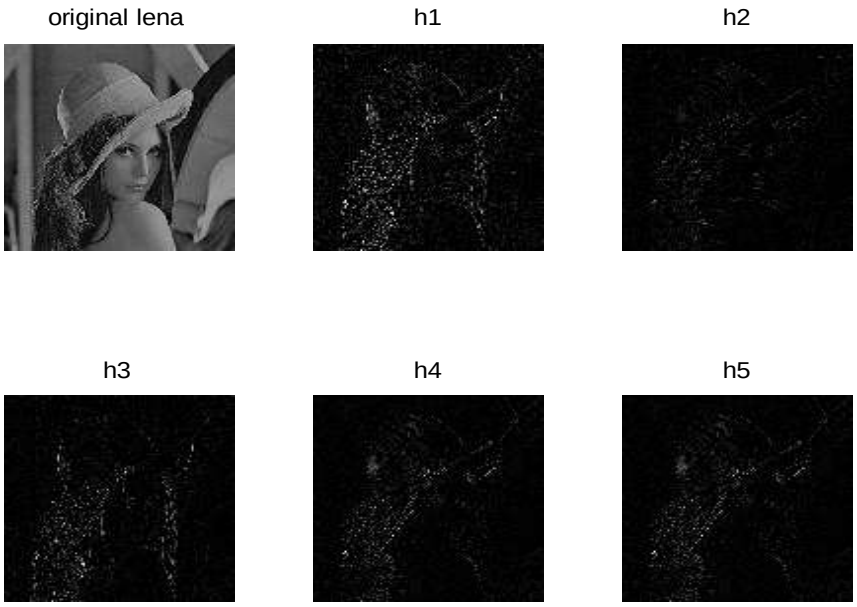


Figure [IV.5]: Détection des contours.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                   filter operations - smoothing
%                    average, median vs gaussian
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% median filter 4x4
im2 = imnoise(im,'salt & pepper',0.04);
im3 = medfilt2(im2, [4, 4]);

h6 = ones(4,4)./(4*4); % box filter
im4 = imfilter(im2, h6);

h7 = fspecial('gaussian');
im5 = imfilter(im2, h7);

figure;
subplot(2,2,1); imshow(im2); title('noisy lena');
subplot(2,2,2); imshow(im3); title('median filter');
subplot(2,2,3); imshow(im4); title('average');
subplot(2,2,4); imshow(im5); title('gaussian');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%            filter operations - edge detection  LoG
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
figure(8);
sgm = 2; t = [0.2 0.4 0.8 0.95];
for k=1:length(t)
    edge_map=edge(im, 'canny', t(k), sgm);
    subplot(2,2,k); imshow(edge_map);
    str=sprintf('LoG: t=%1.1f', t(k)); title(str);
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%                 filter operations - edge detection
%                sobel, laplacian of gaussian, prewitt
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
figure;
im1 = imread('lena.bmp');
im2 = imread('cameraman.tif');
k1 = fspecial('log');
k2 = fspecial('sobel');
k3 = fspecial('prewitt');
subplot(2,3,1); imshow(imfilter(im1, k1)); title('LoG');
subplot(2,3,2); imshow(imfilter(im1, k2));  title('Sobel');
subplot(2,3,3); imshow(imfilter(im1, k3));  title('Prewitt');
subplot(2,3,4); imshow(imfilter(im2, k1));  title('LoG');
subplot(2,3,5); imshow(imfilter(im2, k2));   title('Sobel');
subplot(2,3,6); imshow(imfilter(im2, k3)); title('Prewitt');


return;
```



Figure [IV.6]: Détection des bruits.

La figure suivant présente une image selon l'operateur de filtrage LOG.

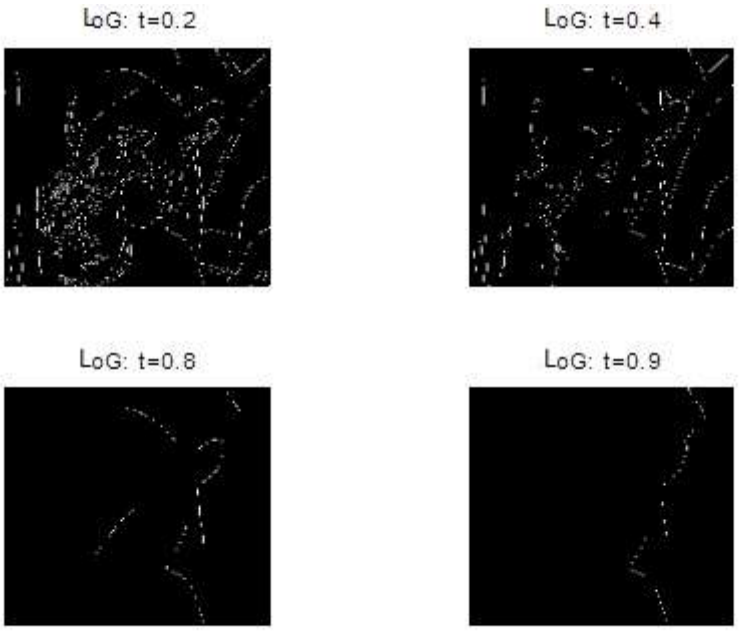


Figure [IV.7]: Application du filtre LOG à la détection de contours.

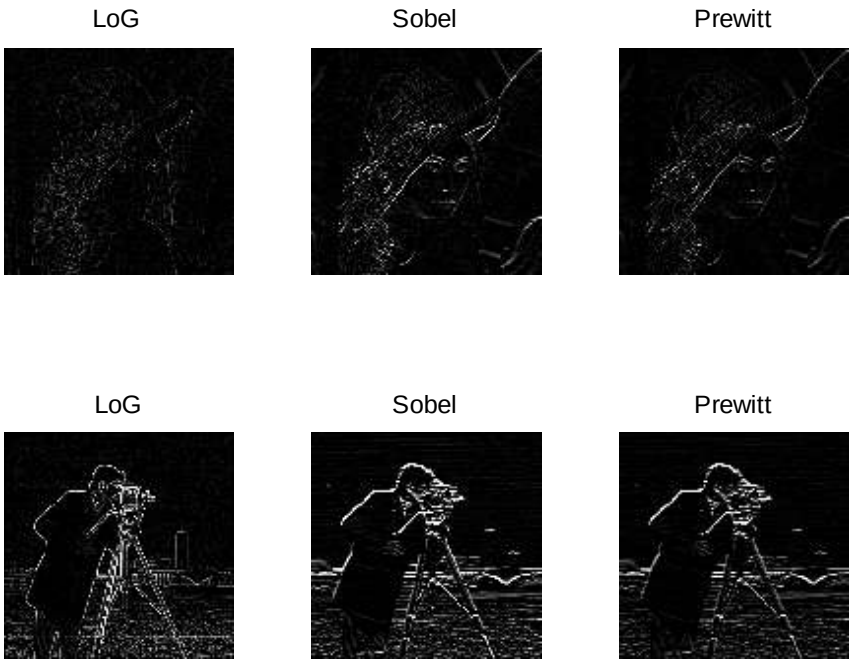


Figure [IV.8] : Détection des contours selon les opérateurs Log, Sobel et Prewitt.

IV.4.La reconnaissance de forme :

On désigne par reconnaissance de formes (ou parfois reconnaissance de motifs) un ensemble de techniques et méthodes visant à identifier des motifs à partir de données brutes afin de prendre une décision dépendant de la catégorie attribuée à ce motif. On considère que c'est une branche de l'intelligence artificielle qui fait largement appel aux techniques d'apprentissage automatique et aux statistiques.

Les formes ou motifs à reconnaître peuvent être de nature très variée. Il peut s'agir de contenu visuel (code barre, visage, empreinte digitale...) ou sonore (reconnaissance de parole), d'images médicales (rayon X, EEG, IRM...) ou multi spectrales (images satellitaires) et bien d'autres [20].

La reconnaissance de formes est utilisée dans plusieurs domaines d'activités. Parmi ces domaines on peut citer :

- Recherche d'images par le contenu : Cette technologie est actuellement intéressante pour la recherche de données sur l'imagerie médicale, ou cartographiques.

- Classification de documents : Son activité est essentielle dans de nombreux domaines économiques : elle permet d'organiser des corpus documentaires, de les trier, et d'aider à les exploiter dans des secteurs tels que l'administration, l'aéronautique, la recherche sur internet, les sciences.
- Reconnaissance de l'écriture manuscrite: Cette technologie fait appel à la reconnaissance de forme, mais également au traitement automatique du langage naturel. Cela veut dire que le système, tout comme le cerveau humain, reconnaît des mots et des phrases existant dans un langage connu plutôt qu'une succession de caractères. Ceci améliore grandement la robustesse.

IV.5.Analyse en Composantes Principales (ACP):

L'analyse en composantes principales (ACP) [17] est une méthode mathématique qui peut être utilisé pour simplifier un ensemble de données, en réduisant sa dimension. Elle est utilisée pour représenter efficacement les images de visages, qui peuvent être approximativement reconstruites à partir d'un petit ensemble de poids et d'une image de visage standard (*Eigenpicture*). Ces poids sont obtenus en projetant l'image dans un espace de visage engendré par les visages propres (*Eigenfaces*). Ainsi au lieu de stocker une vectrice image de taille Résolution R, on stocke un vecteur de taille N. Dans l'approche ACP la normalisation d'éclairage est toujours indispensable [6].

- **La reconnaissance de formes dans des espaces de grande dimension [17] :**
  - ✓ Les problèmes surviennent lors de l'exécution reconnaissance dans un espace de grande dimension. (Par exemple, la malédiction de la dimensionnalité).
  - ✓ Des améliorations significatives peuvent être réalisées en première cartographie des données dans un espace inférieur-dimensionnalité.

$$x = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_N \end{bmatrix} --> réduire\ la\ dimensionnalité --> y = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_K \end{bmatrix} \ (K << N)$$

  - ✓ L'objectif d'ACP est de réduire la dimensionnalité des données tout en conservant comme autant que possible de la variation présente dans le jeu de données d'origine.

- **La réduction de la dimensionnalité [17] :**
  - ✓ PCA permet de calculer une transformation linéaire qui fait correspondre les données d'une grande espace de dimension à un espace de dimension inférieure.

$$b_1 = t_{11}a_1 + t_{12}a_2 + \ldots + t_{1n}a_N$$

$$b_2 = t_{21}a_1 + t_{22}a_2 + \ldots + t_{2n}a_N$$

$$b_k = t_{K1}a_1 + t_{K2}a_2 + \ldots + t_{KN}a_N$$

$$\text{Ou } y = Tx \text{ où } T = \begin{bmatrix} t_{11} & t_{12} & \ldots & t_{1N} \\ t_{21} & t_{22} & \ldots & t_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ t_{K1} & t_{K2} & \ldots & t_{KN} \end{bmatrix}$$

- **Base de moindre dimension [17] :**
  - ✓ Rapprochant les vecteurs de trouver une base dans une moindre dimension appropriée l'espace.
    1. de dimensions supérieures représentation de l'espace:

$$x = a_1v_1 + a_2v_2 + \cdots + a_Nv_N$$

$v_1, v_2, \ldots, v_N$ Est une base de l'espace à N dimensions.

2. Basse-dimensionnelle représentation de l'espace:

$$\hat{x} = b_1u_1 + b_2u_2 + \cdots + b_Ku_K$$

$u_1, u_2, \ldots, u_K$ Est une base de l'espace à k dimensions.

Remarque: Si les deux bases ont la même taille$(N = k)$, donc $x = \hat{x}$

- **La perte de l'information** :
  - ✓ Réduction de dimension implique la perte d'informations!
  - ✓ Préserver autant d'information que possible, c'est-à-dire Minimiser $\|x - \hat{x}\|$
- **Comment déterminer l'espace de dimension plus faible?**

  Le meilleur espace de petite dimension peut être déterminé par les "meilleurs" vecteurs propres de la matrice de covariance de x (par exemple, les vecteurs propres

correspondant aux valeurs propres les plus grandes - Aussi appelés «composantes principales»).

- **Méthodologie :**

Supposons que $x_1, x_2, \dots, x_M$ sont Nx1 vecteurs

Etape 1 : $\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$

Étape 2: soustraire la moyenne $\Phi_i = x_i - \bar{x}$

Étape 3: former la matrice $A = [\Phi_1 \Phi_2 \dots \Phi_M]$ (Matrice N x M), puis calculer:

$$C = \frac{1}{M}\sum_{n=1}^{M} \Phi_n \Phi_n^T = AA^T \qquad [IV.1]$$

(Matrice de covariance de l'échantillon, NxN, caractérise la dispersion des données)

Etape 4: calculer les valeurs propres de C :$\lambda_1 > \lambda_2 > \cdots > \lambda_N$

Etape 5: calculer les vecteurs propres de C :$u_1, u_2, \dots, u_N$

- ✓ Puisque C : est symétrique $u_1, u_2, \dots, u_N$ forme une base, (c'est à dire, tout vecteur x ou réellement$(x - \bar{x})$, peut s'écrire comme une combinaison linéaire des vecteurs propres):

$$x - \bar{x} = b_1 u_1 + b_2 u_2 + \cdots + b_N u_N = \sum_{i=0}^{N} b_1 u_1 \qquad [IV.2]$$

Etape 6: (étape de réduction de dimension) ne retenir que les termes correspondants aux k plus grandes valeurs propres:

$$\hat{x} - \bar{x} = \sum_{i=1}^{K} b_i u_i \text{ Où } k << N \qquad [IV.3]$$

- ✓ La représentation de $\hat{x} - \bar{x}$ dans la base $u_1, u_2, \dots, u_K$ n'est donc

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}$$

- **transformation linéaire implicite par l'APC :**

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix} = \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_K^T \end{bmatrix} (x - \bar{x}) = U^T (x - \bar{x})$$

- **Interprétation géométrique :**

✓ ACP projette les données selon les directions où les données varient le plus.

✓ Ces directions sont déterminées par les vecteurs propres de la matrice de covariance correspondant aux plus grandes valeurs propres.

✓ L'amplitude des valeurs propres correspondant à la variance des données selon les directions des vecteurs propres.

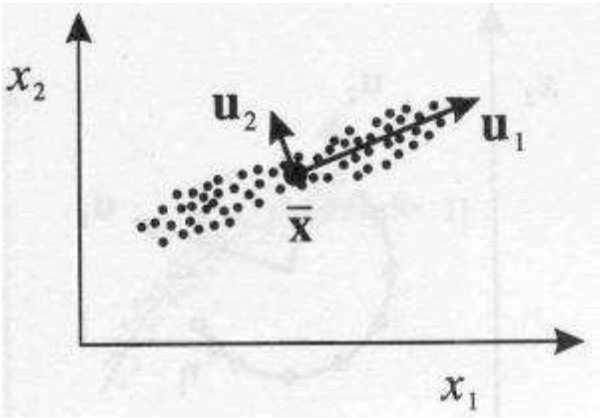


Figure [IV.9] : Interprétation géométrique [17].

- **Les propriétés et les hypothèses d'ACP :**

✓ Les nouvelles variables (c-à-d b_i's) ne sont pas corrélés.

La covariance de b_i's est:

$$U^T C U = \begin{bmatrix} \lambda_i & 0 & 0 \\ 0 & \lambda_2 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \lambda_K \end{bmatrix}$$

✓ La matrice de covariance ne représente que les statistiques de deuxième ordre parmi la vectrice valeur.

✓ Depuis les nouvelles variables sont des combinaisons linéaires des variables d'origine, il est habituellement difficile à interpréter leur signification.

- **Comment choisir les composants principaux?**
  - ✓ Pour choisir k, utilisez les critères suivants:

$$\frac{\sum_{i=1}^{N}\lambda_i}{\sum_{i=1}^{K}\lambda_i} > \text{Seuil (par exemple, 0,9 ou 0,95)}$$

- **Quelle est l'erreur due à la réduction de la dimensionnalité?**
  - ✓ Nous avons vu plus haut qu'un vecteur x original peut être construit à l'aide de ses principales composantes:

$$\hat{x} - \bar{x} = \sum_{i=1}^{K} b_i u_i \text{ ou } \hat{x} = \sum_{i=1}^{K} b_i u_i + \bar{x} \qquad [IV.4]$$

  - ✓ On peut montrer que la base faible dimension basée sur des composants principaux minimise l'erreur de reconstruction:

$$e = \|x - \hat{x}\|$$

  - ✓ On peut montrer que l'erreur est égale à:

$$e = \frac{1}{2}\sum_{i=K+1}^{N} \lambda_i \qquad [IV.5]$$

- **Standardisation :**
  - ✓ Les principaux composants sont dépendants les unités utilisées pour mesurer l'original
    les variables ainsi que sur la gamme des valeurs qu'ils assument.
  - ✓ Nous devrions toujours normaliser les données avant d'utiliser l'ACP.
  - ✓ Une méthode de normalisation commun est de transformer toutes les données aient zéro moyenne et écart-type de l'unité

$\frac{x_i - \mu}{\sigma}$ (μ et σ : La moyenne et l'écart de x_i)

- **ACP et à la classification :**
  - ✓ ACP n'est pas toujours une procédure dimensionnalité de réduction optimale pour la classification objective:

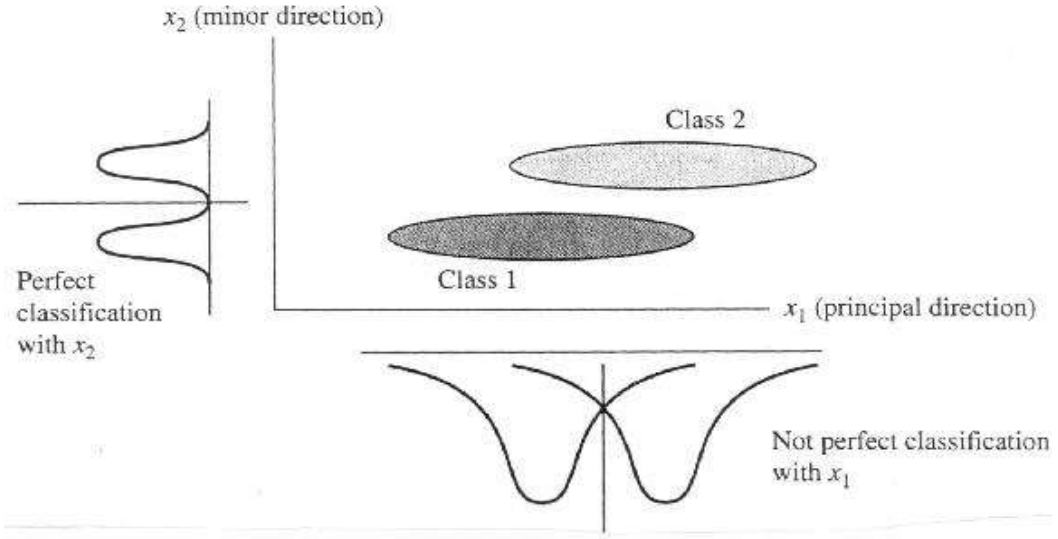


Figure [IV.10] : Les classifications d'ACP [17].

ACP est une technique rapide, simple et populaire dans l'identification de modèle, c'est l'une des meilleures techniques. Les projections d'ACP sont optimales pour la reconstruction d'une base de dimension réduite. Cependant, ACP n'est pas optimisé pour la séparabilité (discrimination) de classe. Une alternative qui est l'analyse discriminante linéaire ADL tient compte de ceci.

IV.6.Analyse Discriminante Linéaire :

L'analyse discriminante linéaire (ADL) [18] et les discriminants linéaires de Fisher (DLF) relatifs, sont utilisés pour trouver la combinaison linéaire des caractéristiques qui séparent le mieux les classes d'objet ou d'événement. Les combinaisons résultantes peuvent être employées comme classificateur linéaire, ou généralement dans la réduction de caractéristiques avant la classification postérieure. ADL est étroitement lié à ACP, du fait que tous les deux recherchent des combinaisons linéaires des variables qui représentent au mieux les données. ADL essaye explicitement de modeler la différence entre les classes des données. ACP quant à elle, ne tient pas compte des différences entre les classes.

- **Moyenne d'images [18]:**
  - ✓ Soit $X_1, X_2, \dots, X_n$les classes de visage dans la base de données et que chaque classe visage X_i, i = 1,2, ..., n et k facial images x_j, j = 1,2, ..., k.
  - ✓ Nous calculons l'image moyenne$\mu_i$de chaque classe X_i :

$$\mu_i = \frac{1}{k}\sum_{i=1}^{k} x_j \qquad [IV.6]$$

Maintenant, l'image μ moyenne de toutes les classes de la base de données peut être calculée comme suit:

$$\mu = \frac{1}{n}\sum_{1=1}^{n} \mu_i \qquad [IV.7]$$

- **Matrices de dispersion [18] :**
  - ✓ Nous calculons matrice de dispersion intra-classe:

$$S_W = \sum_{i=1}^{n} \sum_{x_k \in X_i} (x_i - \mu_i)(\mu_i - \mu)^T \qquad [IV.8]$$

  - ✓ Nous calculons la matrice de dispersion inter -classe:

$$S_B = \sum_{i=1}^{n} N_i(\mu_i - \mu)(\mu_i - \mu)^T \qquad [IV.9]$$

- **Projection [18]:**
  - ✓ Nous trouvons que le produit de SW-1 et SB, puis calculer les vecteurs propres de ce produit($S_W^{-1} \times S_B$) , Après avoir réduit la dimension de l'espace des caractéristiques.
  - ✓ Utilisez la même technique qu'*eigenfaces* approchent à réduire la dimension de la matrice de dispersion pour calculer les vecteurs propres.
  - ✓ Formez un matrice U qui représente tous les vecteurs propres de ($S_W^{-1} \times S_B$) en plaçant chaque $U_i$vecteur propre de chaque colonne de cette matrice**.**

$$U = \begin{bmatrix} -0.2621 & 0.3256 & -1.3511 \\ -0.1137 & 0.9922 & 1.0014 \\ 0.6015 & -0.5080 & -5.4206 \end{bmatrix} \begin{matrix} \\ \\ \\ u_1 \quad u_2 \quad u_3 \end{matrix}$$

Chaque face l'image $x_j \in X_i$ peut-être projetée dans cet espace du visage par l'opération

$$\Omega_i = U_T(x_j - \mu) \qquad [IV.10]$$

Donc, LDA est une amélioration de PCA, Construire un sous-espace discriminant qui minimise la dispersion entre les images de même catégorie et maximise la dispersion entre les différentes images de la classe. L'analyse discriminante linéaire est une approche statistique d'appréciation du risque. Son caractère objectif, précis et général fait d'elle un outil de gestion très performant.

IV.7.Bases de Données de Visages :

Plusieurs bases de données contenants des informations qui permettent l'évaluation des systèmes de reconnaissances de visages sont disponibles sur le marché. Toutefois, ces bases de données sont généralement adaptées aux besoins de quelque algorithme spécifiques de reconnaissance, chacune d'elle a été construire avec des conditions d'acquisition d'image de visages diverses (changements d'illumination, de pose, d'expressions faciales) ainsi que le nombre de sessions pour chaque individu.

A. ORL 1994 [10]:

Notre base de données de visages, (anciennement 'La base de données ORL de Faces'), contient un ensemble d'images de visages prises entre Avril 1992 et Avril 1994 [10].

Cette base contient des images de visage de 40 personnes, avec 10 images pour chacune, pour la plupart des sujets, les 10 images ont été tirées pour différentes heures et avec différentes conditions d'éclairage, mais toujours sur un font foncé. La taille d'image est 92x112, Les fichiers sont au format PGM [10].

Les limitations de cette base de données sont : le nombre limité de personne, la non – conformité des conditions d'éclairage d'une image à une image, et a non assignation de

toutes les images par les différentes expressions faciales, rotation de la tête, ou conditions d'éclairage [13].

B. IFD 2002 [15] :

Cette base de données contient des images de 40 sujets distincts avec 11 poses différentes pour chaque individuel. Si disponibles, quelques images additionnelles sont également incluses pour quelques individus. Toutes les images ont un fond homogène lumineux et les sujets sont en position droite et frontale. Pour chaque individuel, nous avons inclus la pose suivante pour le visage : regardant l'avant, regardant à gauche, regardant la droite, cherchant, cherchant vers la gauche, cherchant vers la droite, regardant vers le bas. En plus de la variation de la pose, des images avec quatre émotions neutre, sourire, le rire,…, sont également incluses pour chaque individu. Comme exemple, des images correspondant à un individu sont montrées ci-dessous.



Figure [IV.11] : Images d'une personne dans la base IFD [15].

Les dossiers sont dans le format de JPEG. La taille de chaque image est les Pixel 640x480, avec 256 niveaux gris par Pixel. Les images sont organisées dans deux annuaires principaux - des mâles et des femelles. Dans chacun de ces annuaires, il y a des sous-répertoires avec des noms comme nombres, où chaque index correspondant à un individu. Dans chacun de ces annuaires, il y a 11 images différentes de ce sujet, qui ont des noms de la forme N.JPG, où n est le nombre d'image pour ce sujet [15].

IV.8.Analyse Expérimental des Résultats :

Après avoir survolé la littérature de data mining et de quelques méthodes de filtrage ainsi les opérateurs de gradient, on va passer à une étape d'évaluation de deux techniques, afin d'aboutir à la méthode la plus optimale et la plus adéquate. Dans cette partie, on présente les résultats de l'ensemble des tests qu'on a effectué sur les deux bases de données, et en utilisant différentes méthodes et sous différentes conditions.

L'évaluation est effectuée en utilisant l'expert d'identification de visage. C'est une source ouverte MATLAB (R 2010 a), et un PC de processus (i5), fréquence (2.67 GHZ) et une RAM (6 G Byte) sur un système d'exploitation windows 8 de (64 bites).

La Comparaison est faite sur la base du taux d'exactitude d'identification. Les résultats comparatifs obtenus en examinant des deux algorithmes ACP et ADL sur l'IFD base de données indienne de visage et d'ATT 60% pour le temps d'apprentissage et 40% pour le temps de teste. La base de données a été utilisée pour la comparaison de l'exécution. Ces deux bases de données ont été choisies parce que l'ATT contient les changements très petits d'image commutateur de l'orientation des images pour chaque sujet impliqué, tandis que l'IFD contient un ensemble de 10 images pour chaque sujet où chaque image est orientée dans un angle différent comparé à des autres.

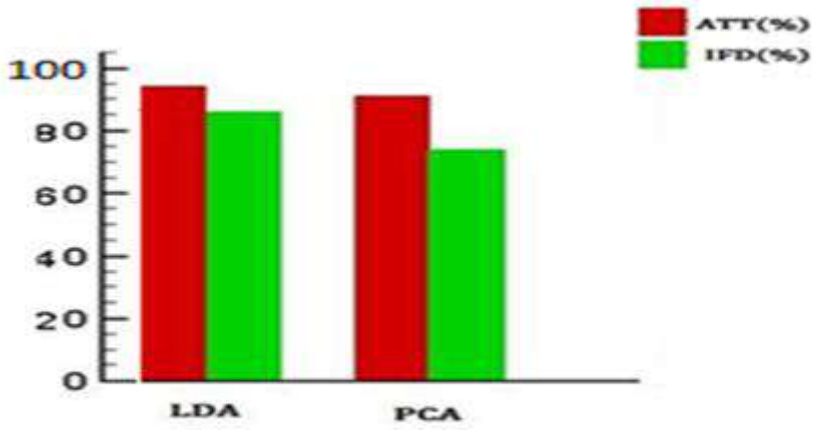


Figure [IV.12] : Le pourcentage de reconnaissance pour les algorithmes ACP et ADL.

- Au-dessus de l'analyse montre l'exécution des deux algorithmes d'ACP et ADL sur la base de données de l'ATT et de l'IFD. Les déférents points que nous avons observés dans cette expérience :
  - ✓ On observe que les deux algorithmes utilisés sur la base de données ATT et IFD.
  - ✓ ADL a le taux d'exactitude le plus élevé d'identification qu'ACP.
  - ✓ On observe que le taux d'identification de la base de données d'ATT est plus haut comme comparent à la base de données d'IFD. Dans cette base de données, chaque sujet est dépeint avec des angles fortement variables d'orientation. Il prouve également que chaque image a le fond riche région que la base de données d'ATT.
- Après cette étude qui présente le pourcentage de reconnaissance de visage pour les algorithmes ACP et ADL sur la base ATT et IFD. On va voir le temps de traitement en ms/img pour les mêmes algorithmes et la base de données sur la courbe suivant :

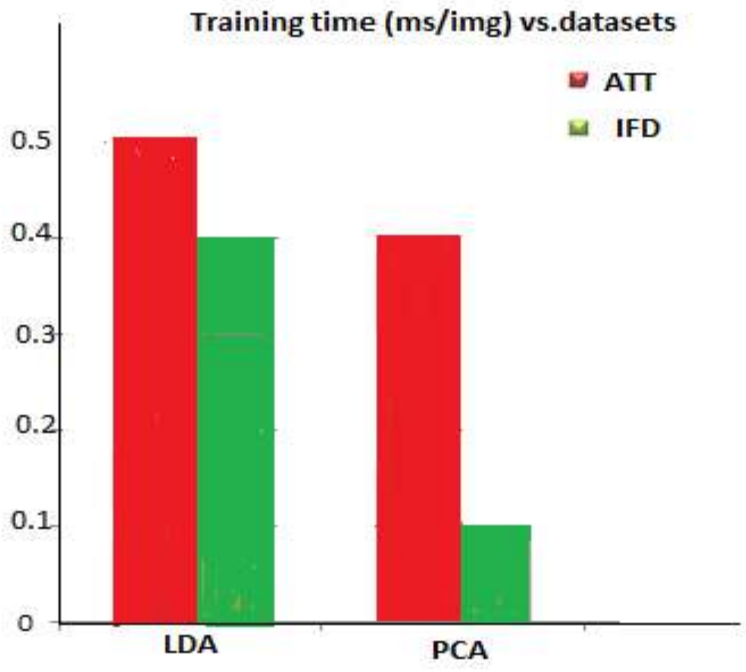


Figure [IV.13]:Temps de traitement pour les algorithmes de ACP et ADL sur la base ATT et IFD.

- Donc les déférents points que nous avons observés :
  - ✓ On remarque que le temps d'apprentissage pour l'algorithme ADL et plus que l'algorithme ACP.

- ✓ Ainsi qu'on a vu la déférence dans les algorithmes ACP et ADL est très petits dans la base de données ATT contrairement dans IFD.

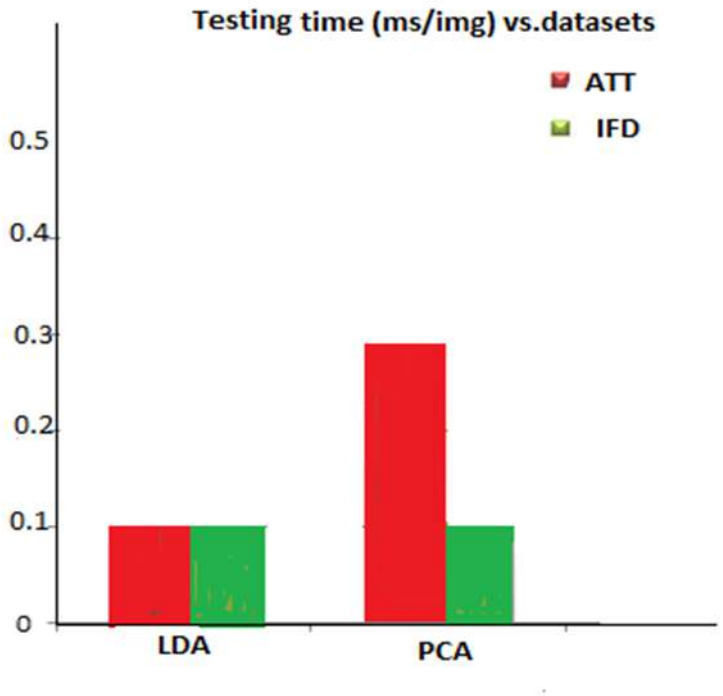


Figure [IV.14] : Temps de test pour les algorithmes ACP et ADL sur la base ATT et IFD.

Selon la courbe précédente qui représente le temps de test des deux algorithmes ACP et ADL.

- ✓ Le temps de test de l'algorithme ACP plus élevé pour la base de donnée ATT et très petits pour IFD.
- ✓ Le temps de test de l'algorithme ADL est presque le même pour la base ATT et IFD.
- ✓ Ainsi on remarque que ACP prend un temps de test plus élevé par rapport que ADL pour la base ATT, mais pour la base de données IFD le temps de test presque le même.

IV.9.Conclusion :

Nous avons utilisé deux algorithmes de l'identification ACP et ADL. La performance de visage a été calculée en termes d'exactitude d'identification. On l'observe que le taux d'identification de la base de données ATT est plus haut comme comparent à la base de données d'IFD. On observe que les deux algorithmes sont utilisés sur la base de données, et ACP précèdent algorithmes d'ADL.

Conclusion générale.

Depuis leur création, les méthodes d'exploration des données ont été étudiées par de nombreux chercheurs et ont été adaptées à de nombreuses situations. Cependant, il reste encore beaucoup de choses à développer.

On sait que la plupart de ces méthodes ont meilleure application dans le cas où il y a plus de prédicteurs que d'observations et qu'il y a une forte colinéarité entre ces prédicteurs, sachant qu'il est toujours plus intéressant de pouvoir étudier plus précisément la qualité d'un modèle lorsqu'il a été construit.

Le *data mining* est l'extraction d'informations prédictives cachés dans de grandes base de données. C'est une technologie nouvelle et puissante qui donne la possibilité aux entreprises de se concentrer sur les informations les plus importantes dans leurs data *warehouses*. Les outils du *data mining* peuvent prédire les futurs tendances et actions, permettant de prendre les bonnes décisions. C'est ce qui rend le *data mining* la technologie la plus importantes.

L'analyse en composantes principales est une excellente approche pour explorer la structure des données en tenant compte de leur caractère multidimensionnel. Cette méthode est l'origine de la plupart des méthodes descriptives multidimensionnelles comme l'analyse discriminante linéaire, L'analyse en composantes principales sert à mieux connaitre les données sur lesquelles on travaille en construisant de nouvelles variables artificielles et fournit des représentations graphiques qui permettent de visualiser des liaisons entre les variables.

La reconnaissance des visages par l'approche *eigenfaces*, est très utilisée pour des raisons de sécurité, utilisée par les institutions policières, agents d'immigration, dans les prisons et pour permettre à une personne l'accès à une zone sécurisée.

D'après les points que nous avons cités dans ce cadre de ce thème, nous pouvons juger le domaine *data mining* comme un domaine apte à l'évolution et riche en sujets de recherche qui attend, celui qui vient pour relever ses défis.

Bibliographies

[1] Cours des détections des contours
http://www.optique-ingenieur.org/fr/cours/OPI_fr_M04_C05/co/Grain_OPI_fr_M04_C05.html

[2] BELMEKKI M., ZABOUBI A., *contribution de la base de données d'un logiciel de formation de portraits robots*, mémoire de fin d'étude pour l'obtention du diplôme d'ingénieur d'état, 2004.

[3] SELLAMI L., *Approche Data Mining pour la Détection d'Intrusions*, Mémoire de Magistère, Université Abderrahmane Mira de Bejaia.

[4] CHAMI D., *Une plateforme orientée agent pour le data mining*, mémoire de Magister, Université HADJ LAKHDAR – BATNA ,2009-2010.

[5] L'encyclopédie en ligne WIKEPEDIA 2009.

[6] JOURANI R., *Reconnaissance de visages*, thèse pour l'obtention du « Diplôme des Etudes Supérieures Approfondies », Université Mohammed V-Agdal rabat, 2006, [http://www.adultpdf.com/].

[7] La détection automatique de clusters, Rapporté de
http://www.eisti.fr/~lassi/PFE/ID/DataMining/SITE/detection.htm.

[8] ZENASNI H., *Projet de génome I : data mining sur méthode du pseudo potentiel empirique*, mémoire de magister, université Abou-BAKR BELKAID de Tlemcen, 2007-2008.

[9] Détection de contours par approximation locale par splines ,
http://www.optiqueingenieur.org/fr/cours/OPI_fr_M04_C05/co/Grain_OPI_fr_M04_C05_4.html.

[10] The Database of Faces,
http://www.cl.cam.ac.uk/Research/DTG/attarchive/pub/data/att_faces.tar .

[11] CHAOUCHI CH., *Etude et implémentation* de *détecteurs de contours*, Mémoire de fin d'études pour l'obtention du diplôme d'ingénieur d'état en Electronique, Ecole Nationale Polytechnique ,2004.

[12] PIATETSKY-SHAPIRO G., *Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from «university» to «business» and «analytics», Data mining and Knowledge Discovery, 15(1), 99-105.*

[13] BOUDJELLAL S., *Détection et identification de personne par méthode biométrique*, mémoire de magister, université mouloud MAMMERI TIZI-OUZOU (UMMTO).

[14] SAIDI F., *Génome des matériaux II : Conception de revêtements ultra durs sous forme de Super-réseaux en utilisant l'approche du Datamining*, mémoire de magister, université Abou-BAKR BELKAID de Tlemcen., 2008.

[15] The Indian Face Database.
http://vis-www.cs.umass.edu/~vidit/IndianFaceDatabase/

[16] GUETTAL L., BEDOUI L., *Authentification de visages par la méthode d'analyse discriminante linéaire de Fischer,* Mémoire de fin d'études pour l'obtention du Diplôme d'Ingénieur d'état, université Mohamed KHEIDER de Biskra, 2008.

[17] Principal Components Analysis (PCA), Lecture,
http://www.cse.unr.edu/~bebis/MathMethods/PCA/lecture.pdf

[18] Buddharaju P., *Face Recognition using PCA (Eigenfaces) and LDA (Fisherfaces)*, Slides for Lectures; http://www1.cs.columbia.edu/.../2010/eigenfisherfaces.ppt

[19] SALMI A., *Reconnaissance des visages par l'ACP (Concept d'Eigenfaces),* cours ACP de MESFIOUI Met de MEUNIER F., université du québec à trois –rivières ,2011.

[20] POISSON D., MAHJOUB S., *La reconnaissance de Forme : Comment améliorer les techniques de reconnaissance de forme 3D ?,* Mémoire de Recherche, Paris, 2010.

[21] Dreyfus G., Martinez J-M., Samuelides M.,
Gordon B., Badran F., Thiria S., *Apprentissage statistique Réseaux de neurones ; Cartes topologiques Machines à vecteurs supports,* livre : *Cet ouvrage est la troisième édition, avec mise à jour et nouveau titre, de l'ouvrage paru à l'origine sous le titre « Réseaux de neurones – Méthodologie et applications » (ISBN : 978-2-212-11464-5).*