République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Batna 2 – Mostefa Ben Boulaïd
Faculté de Technologie
Département de l'électronique

Thèse

Préparée au sein du Laboratoire d'Automatique Avancée et d'Analyse des Systèmes (LAAAS)

Présentée pour l'obtention du diplôme de :

Doctorat en Sciences
Option : Traitement du Signal

Sous le Thème :

Reconnaissance d'activités humaines en utilisant les descripteurs spatio-temporels 2D/3D

Présentée par :

KHELALEF Aziz

Devant le jury composé de :

M. LOUCHENE Ahmed	Prof.	Université Batna 2	Président
M. BENOUDJIT Nabil	Prof.	Université Batna 2	Rapporteur
M. ABABSA Fakhreddine	Prof.	Arts et Métiers ParisTech	Co-Rapporteur
M. GOLEA Noureddine	Prof.	Université Oum El Bouaghi	Examinateur
M. KAZAR Okba	Prof.	Université Biskra	Examinateur
Mme. MOUSS Leila Hayet	Prof.	Université Batna 2	Examinateur

Juin 2020

Je Dédie ce travail :

A mon père et ma mère.
A ma femme et mes petites filles Nehal et Ines.
A mes frères et sœurs.
A ma famille.
A tous ceux qui m'aiment et ceux que j'aime.

Remerciements.

Nous rendons grâce à Dieu qui nous a donné l'aide, la patience et le courage pour accomplir ce travail.

Je tiens à exprimer un immense merci et une reconnaissance éternelle envers le Professeur **BENOUDJIT Nabil** d'avoir accepté de diriger mes travaux de thèse, de son encouragement et pour les recommandations qu'il m'a prodiguées et qui m'ont été d'un grand apport lors de la réalisation de ce travail.

Un grand merci aussi au Professeur **ABABSA Fakhreddine**, co-directeur de thèse. D'avoir accepté de travailler avec nous dans ce travail, pour les recommandations, et les précieuses suggestions.

Sans oublier le Professeur **MELGANI Farid**, pour les recommandations ainsi que son accueil à Trento dans le cadre du stage BigPro, et les moyens qu'il m'a offerts pour réaliser le dernier chapitre.

Je remercie le Professeur **LOUCHENE Ahmed** qui m'a honoré par sa présence en qualité de président de jury.

Je tiens à adresser mes plus vifs remerciements aux membres de jury : Professeur **GOLEA Noureddine**, Professeur **KAZAR** Okba et le Professeur **MOUSS Leila Hayet** pour avoir accepté de juger mon travail.

Je remercie aussi tous ceux qui ont contribué à l'élaboration de ce travail de près ou de loin et qui méritent d'y trouver leurs noms.

Khelalef Aziz

Résumé.

Dans cette thèse, nous nous sommes intéressés principalement au problème de la reconnaissance d'activités humaines en utilisant les descripteurs spatio-temporels 2D/3D. Ce domaine est un axe de recherche très actif dont le but est de doter les machines du pouvoir d'analyse et d'interprétation des mouvements réalisés par des humains dans une séquence vidéo.

Nous avons présenté le domaine de reconnaissance d'activités humaines, les problématiques posées, ainsi que les solutions présentées dans la littérature. De plus, nous avons réalisé une étude détaillée sur les réseaux de neurones artificiels et l'apprentissage profond (*deep learning*) ainsi que les techniques proposées dans l'état de l'art basées sur ce dernier.

Nous avons contribué dans ce domaine par la proposition de quatre techniques de reconnaissance d'activités humaines. La première technique présentée pour la reconnaissance des activités dans les séquences vidéo est basée sur l'extraction des skeletons et la transformée en cosinus discrète DCT pour l'extraction des caractéristiques et la SVM pour la classification des activités. Nous avons présenté aussi une variante de cette technique pour la reconnaissance des activités image par image en temps réel, cette dernière est basée sur l'extraction des silhouettes et la DCT pour l'extraction des caractéristiques et les réseaux de neurones artificiels RBF pour la classification des activités. Les deux techniques proposées ont donné des résultats très satisfaisants et très performants, cependant, elles présentent un point faible commun qui est l'algorithme de classification.

Pour pallier à ce problème, nous avons proposé une nouvelle méthode basée sur un nouveau descripteur appelé BSTM et les réseaux de neurones à convolution CNN pour la reconnaissance. Cette approche a donné des résultats très performants, en surpassant les techniques conventionnelles et en donnant des résultats comparables aux nouvelles techniques basées sur l'apprentissage profond.

Malgré cela, notre méthode s'est révélée incapable de faire la reconnaissance des activités image par image en temps réel à cause de la technique d'extraction du descripteur BSTM, pour cela, nous avons proposé une quatrième technique de reconnaissance d'activités humaines totalement automatisée en utilisant l'apprentissage par transfert du modèle YOLO

(*You Only Look Once*) conçu à la base pour la reconnaissance d'objet. Nous avons proposé aussi un protocole de fusion pour adapter notre technique pour la reconnaissance des activités dans les séquences vidéo.

Les résultats expérimentaux ont montré que l'approche proposée a donné des résultats très encourageants lors de la reconnaissance des activités image par image, et a surpassé toutes les techniques présentées dans la littérature lors de l'utilisation du protocole de fusion pour la reconnaissance des activités dans les séquences vidéo.

Abstract.

In this thesis, we are mainly interested in the problem of human activity recognition using 2D/3D spatio-temporal descriptors. This is a very active area of research, in which the aim is to equip machines with the power of analysis and interpretation of the movements made by humans in a video sequence.

We presented the field of human activity recognition, the main issues, as well as the solutions presented in the literature. In addition, we have carried out a detailed study on artificial neural networks and deep learning, we also presented the main techniques proposed in the state of the art based on it.

We have contributed in this area by proposing four techniques for human activity recognition. First, we proposed a new technique for human activity recognition is video sequences, it is based on skeletons and the discrete casinos transform DCT to extract the descriptors, and the SVM for the classification of activities. We also presented a variant of this technique for human activity recognition frame by frame in real time, it is based on silhouettes and DCT for characteristics extraction and RBF artificial neural networks for classification. The two proposed techniques have given very satisfied and very efficient results; however, they have a common weak point which is the classification algorithm.

To overcome this problem, we have proposed a third method based on a new descriptor called BSTM and CNN convolutional neural networks for recognition. This approach has given very effective results, surpassing conventional techniques and giving comparable results against new techniques based on deep learning.

Despite this, our method has proved incapable of recognizing frame-by-frame activities in real time because of the BSTM descriptor extraction technique, for this we have proposed a fourth technique for human activity recognition completely automated, using transfer learning of the YOLO model (*You Only Look Once*) originally designed for object recognition. We also proposed a fusion protocol to adapt our technique for human activity recognition in video sequences.

Experimental results have shown that the proposed approach has given very encouraging results when recognizing frame-by-frame activities, and has surpassed all the techniques presented in the literature when using the fusion protocol for human activity recognition in video sequences.

ملخص.

نهتم في هذه الأطروحة بشكل أساسي بمشكلة التعرف على الأنشطة البشرية باستخدام الواصفات المكانية والزمانية. هذا المجال من الأبحاث نشط للغاية، يهدف إلى تزويد الآلات بالقدرة على تحليل وتفسير حركات البشر في الفيديو.

لقد قمنا بالتعريف بهذا المجل، الصعوبات المواجهة فيه، وكذلك الحلول المقدمة، بالإضافة إلى ذلك، قمنا بإجراء دراسة مفصلة عن الشبكات العصبية الاصطناعية والتعلم العميق وكذلك التقنيات المقترحة في هذا المجال على أساس هذا الأخير.

لقد ساهمنا في هذا المجال من خلال اقتراح أربع تقنيات للتعرف على الأنشطة البشرية. تعتمد التقنية الأولى المقدمة على التعرف على الانشطة البشرية في لقطات الفيديو على استخراج الهياكل العظمية والتحويل الى مجال DCT لاستخراج الخصائص وSVM من اجل تصنيف الانشطة البشرية، كما قدمنا ايضا طريقة اخرى من اجل التعرف على النشاطات البشرية مباشرة في كل إطار من الفيديو هده الطريقة تعتمد على استخراج هيئة الانسان (silhouettes) والتحويل DCTمن اجل استخراج المميزات والشبكات العصبية المتعددة الطبقات RBF من اجل التصنيف. أعطت الطريقتان المقترحتان نتائج مرضية وفعالة للغاية، ومع ذلك، فلديهما نقطة ضعف مشتركة وهي خوارزمية التصنيف.

للتغلب على هذه المشكلة، اقترحنا طريقة جديدة تستند إلى واصف جديد يسمى BSTM وCNN الشبكات العصبية العميقة للتعرف عليها. أعطى هذا النهج نتائج فعالة للغاية، وتجاوز التقنيات التقليدية كم اعطى نتائج مماثلة للتقنيات الجديدة القائمة على التعلم العميق.

على الرغم من ذلك، أثبتت هده الطريقة عدم قدرتها على التعرف على الانشطة في الوقت الفعلي بسبب تقنية استخراج واصف BSTM، ولهذا اقترحنا تقنية رابعة للتعرف على الأنشطة البشرية الية بشكل كامل. باستخدام نقل التعلم لنموذج YOLO (You Only Look Once) المصمم أصلاً للتعرف على الاشياء. اقترحنا أيضًا بروتوكول دمج لتكييف أسلوبنا في التعرف على الأنشطة في لقطات الفيديو.

أظهرت النتائج التجريبية أن النهج المقترح قد أعطى نتائج مشجعة للغاية عند التعرف على أنشطة مباشرة إطار تلو الآخر، وقد تجاوز كل التقنيات المقدمة عند استخدام بروتوكول الدمج للتعرف على الأنشطة في لقطات الفيديو.

Sommaire.

Résumé I
Abstract III
ملخص V
Sommaire VI
Liste des figures XI
Liste des tableaux XVII

Introduction Générale

Contexte et problématiques 1
Contributions 5
Plan de la thèse 7

Partie 1 : Etat de l'Art

Chapitre I : Etat de l'Art sur la Reconnaissance d'Activités Humaines.

I.1. Introduction 9
I.2. La reconnaissance d'activités humaines 9
I.3. Terminologie 9
I.4. Le workflow général des techniques de reconnaissance d'activités humaines 10
I.4.1. Pré-traitement et extraction des caractéristiques 10
I.4.2. La classification des activités 10
I.5. Catégorisation des approches de reconnaissance d'activités humaines 11
I.5.1. Approches basées sur les descripteurs globaux 11
I.5.1.1. Approches basées sur les silhouettes 12
I.5.1.1.1. Volume spatio-temporel 12
I.5.1.1.2. L'Image d'Energie du Mouvement (MEI) et l'Image Historique du Mouvement (MHI) 13
I.5.1.2. Approches basées sur le flux optique et le gradient 14
I.5.1.2.1. Flux optique 14
I.5.1.2.2. Histogrammes de Gradients Orientés (HOG) 15
I.5.2. Approches basées sur les descripteurs Locaux 18
I.5.2.1. Descripteur de Harris 19

I.5.2.2. Descripteur SIFT (*Scale Invariant Feature Transform*) 21
I.5.2.3. Descripteur SURF (*Speeded Up Robust Features*) 22
I.5.2.4. Descripteur local de motif binaire LBP (*Local Binary Patterns*)..... 23
I.5.3. Les approches basées sur la modélisation du corps humain 24
I.6. Techniques de Classification 28
I.6.1. Les méthodes de classification supervisées 28
I.6.1.1. K plus proches voisins (KNN) 28
I.6.1.2. Machines à vecteurs de support (SVM) 29
I.6.1.3. Réseaux de neurones 30
I.6.2. Les techniques de classification non supervisées 31
I.6.2.1. K-moyennes (k-means) 32
I.6.2.2. Les Modèles de Markov à états cachés (*Hidden Markov Models* (HMM)) .. 33
I.7. Synthèse 34
I.8. Conclusion 35

Chapitre II : Apprentissage Profond Dans La Reconnaissance d'Activités Humaines.

II.1. Introduction 36
II.2. Réseaux de neurones artificiels 36
II.2.1. Le perceptron (Neurone formel)..... 36
II.2.2. Perceptron Multicouche (*Multi Layer Perceptron* : MLP)..... 38
II.3. L'apprentissage profond (*Deep Learning*) 40
II.3.1. Domaines d'application du *Deep Learning*..... 42
II.3.2. Les réseaux de neurones à convolution (*Convolutional Neural Networks*) 42
II.3.2.1. La couche convolutionnelle 43
II.3.2.2. Fonction d'activation 45
a. ReLU (Unité de Rectification Linéaire) 45
b. Leaky ReLU 45
II.3.2.3. Couche de *Pooling* 46
II.3.2.4. Couche entièrement connectée (*Fully Connected Layer*) 47
II.3.2.5. La couche (*Softmax*)..... 47
II.3.2.6. La couche de classification (*classification layer*) 47
II.3.3. Conception ! de réseaux de neurones à convolutions..... 49

II.3.4. Bases de données utilisées pour l'apprentissage des réseaux CNN 51
II.3.5. Etat de l'art sur les réseaux de neurones à convolution 52
II.4. Les méthodes de reconnaissance d'activités humaines basées sur le Deep Learning 56
II.5. Fusion de l'information temps à travers un réseau de neurones profond CNN 58
II.6. La reconnaissance d'activités humaines en utilisant le descripteur BMI (*Binary Motion Image*) 59
II.7. La reconnaissance d'activités humaines en utilisant l'apprentissage profond séquentiel (*Sequential Deep Learning*) 60
II.8. La reconnaissance d'activités humaines en utilisant les cartes de profondeur et les réseaux de neurones à convolution 62
II.9. La reconnaissance d'activités humaine en exploitant l'apprentissage profond 64
II.10. La classification des vidéos à grande-échelle en utilisant les réseaux de neurones à convolution 65
II.11. Réseau de neurones à convolution à deux canaux pour la reconnaissance d'activités humaines (*Two-Stream*) 66
II.12. Synthèse 68
II.13. Conclusion 70

Partie 2 : Contributions

Chapitre III : Reconnaissance d'Activités Humaines en utilisant la DCT.

III.1. Introduction 71
III.2. Description des bases de données utilisées dans le domaine de la reconnaissance d'activités humaines 71
III.2.1. La base de données de Weizmann 71
III.2.2. La base de données Keck Gesture Dataset 72
III.2.3. La base de données KTH 73
III.3. Critères d'évaluation 74
III.3.1. Taux de reconnaissance 74
III.3.2. Courbe ROC (Roc Curve) 74
III.3.3. Matrice de confusion 76
III.4. Rappel sur la transformée en cosinus discrète (DCT) 77

III.5. Description de la méthode proposée 79
a. Descripteur basé sur les cartes spatio-temporelles et la DCT 79
b. Descripteur basé sur les silhouettes et la DCT 82
III.6. Résultats expérimentaux 82
a. Descripteur basé sur les cartes spatio-temporelles et la DCT 83
a.1. Protocole de test 83
a.2. Discussion des résultats 83
b. Descripteur basé sur les silhouettes et la DCT 87
b.1. Protocole de test 87
b.2. Discussion des résultats 87
III.8. Conclusion 93

Chapitre IV : Reconnaissance d'Activités Humaines en utilisant le Descripteur BSTM et l'Apprentissage Profond

IV.1. Introduction 94
IV.2. Description de la méthode proposée 94
IV.3. Résultats expérimentaux 100
a. La base de données de Weizmann 100
b. Keck Gesture Dataset 104
c. La base de données KTH 107
IV.4. Conclusion 111

Chapitre V : Reconnaissance d'Activités Humaines en utilisant le Modèle YOLO

V.1. Introduction 113
V.2. Présentation du YOLO (*You Only Look Once*) 113
V.3. Architecture du YOLO 114
V.4. Description de la méthode proposée 117
V.5. Résultats expérimentaux 119
V.5.1. Protocole de test 119
V.5.2. Discussion des résultats 119
a. Reconnaissance des activités image par image 119
b. Reconnaissance des activités dans les séquences vidéo 123

c. Application sur la base de données de Weizmann 126
d. Application sur les vidéos YouTube 130
e. Application en temps réel en utilisant la Camera 132
V.6. Conclusion 133
Conclusion générale 135
Production scientifique 139
Références 140

Liste Des Figures.

Introduction Générale

Figure 1 : Exemples d'activités humaines (Cinq images consécutives) tirés de la base de données de Weizmann 2
Figure 2 : Exemple de l'activité « Walk » dans différentes situations..... 3

Chapitre I

Figure I.1 : Le workflow d'une technique de reconnaissance d'activités humaines 10
Figure I.2 : Classification des approches de reconnaissance d'activités humaines 11
Figure I.3 : a) Le volume spatio-temporel, b) La solution de l'équation de poisson, c) Les caractéristiques de saillance, d) Les caractéristiques d'orientation [5] 12
Figure I.4 : Image d'énergie de mouvement (MEI) et l'image de l'historique du mouvement (MHI) [6]. 13
Figure I.5 : a) Image de l'historique du mouvement, b) 2D-DCT par block (8x8). 14
Figure I.6 : a) Image originale, b) Flux optique, c) La séparation de la composante verticale et la composante horizontale, d) Les quatre composantes scalaires, e) Le descripteur final [23]. 15
Figure I.7 : a) Image initiale, b) Luminance de l'image initial, c) Magnitude du gradient par filtrage de Canny, d) Orientation du gradient par filtrage de Canny (les nuances rouges représentent les orientations verticales et les nuances vertes représentent les orientation horizontales) [24]. 16
Figure I.8 : a) Gradient de l'image, b) HOG descripteur (grille =2x2) et 8 orientations, c) PCA-HOG descripteur avec 12 Principal Components, d) Le descripteur HOG reconstituer [8]. 17
Figure I.9 : L'inconvénient du flux optique : sensibilité au changement du fond. 18
Figure I.10 : Diagramme du détecteur de Harris [1]. 20
Figure I.11 : Comparaisons entre le 2D SIFT (à gauche) et le ST SIFT (à droite) [31]. 22
Figure I.12 : Exemple de calcule d'un motif binaire local. 23
Figure I.13 : Le descripteur LBP [36]. 24
Figure I.14 : La méthode l'extraction des caractéristiques LBP proposée dans [37] 24

Figure I.15 : Théorie des points de repère de Johansson [38]. 25
Figure I.16 : Le modèle de squelettes et articulation utilisé par Sheikh et al [9]..... 25
Figure I.17 : La projection de l'activité « Sitting » dans l'espace XYT, les points rouges sont l'action originale, les points Bleus est le modèle reconstruit à partir de la projection sur l'action de base de l'activité « Sitting » [9]..... 26
Figure I.18 : a) Image original (profonde), b) squelettes, c) les coordonnées de référence HOJ3D, d) les coordonnées sphériques projetées [10]..... 26
Figure I.19 : Le modèle 3D construit par Sedai et al [12]. 27
Figure I.20 : Procédure d'extraction de l'Etoile de squelette [11]..... 27
Figure I.21 : Exemple de classification par KNN, classification en 3 classes [40]. 29
Figure I.22 : a) Cas d'un problème linéairement séparable, b) un cas non linéairement séparable (projection des caractéristiques vers un espace linéairement séparable) [40]..... 30
Figure I.23 : Exemple d'un réseau de neurones..... 31
Figure I.24 : Exemple d'une chaine de Markov [40]..... 33

Chapitre II

Figure II.1 : Schéma de fonctionnement d'un perceptron [56]..... 36
Figure II.2 : Fonctions d'activation [58] : a) Heaviside, b) signe, c) Linéaire à seuil, d) Sigmoïde..... 37
Figure II.3 : Fonction d'activation dans un réseau MLP [59]. 38
Figure II.4 : Schéma d'un perceptron Multicouche [56]. 39
Figure II.5 : Algorithme de rétropropagation [59]..... 40
Figure II.6: Machine learning Vs Deep learning. 41
Figure II.7 : Représentations hiérarchiques apprises par un CNN [60]..... 43
Figure II.8 : Exemple de convolution 44
Figure II.9 : Exemples de filtres de convolution : les 96 filtres de la première couche d'AlexNet [61]. 44
Figure II.10 : La fonction ReLU..... 45
Figure II.11 : La fonction Leaky ReLU (α =0.1) 45
Figure II.12 : Exemple d'une couche de convolution avec deux filtres. 46
Figure II.13 : Différentes opérations de *pooling* : à gauche, *Average pooling* et à droite le *Max pooling* (un filtre 2x2, stride =2) [63]. 47
Figure II.14 : Exemple d'un stride=2..... 48

Figure II.15 : Exemple d'un Padding de 1x1. 49
Figure II.16 : Le workflow du fine-tuning. 51
Figure II.17 : Réseaux Vs performances dans l'épreuve Top1 de ImageNet [67]. 52
Figure II.18 : Réseaux Vs paramètres dans l'épreuve Top1 de ImageNet [67]. 52
Figure II.19 : Le premier réseau CNN (LeNet-5) [68]. 53
Figure II.20 : Le réseau d'AlexNet. 54
Figure II.21 : VGG16 [69]. 55
Figure II.22: VGG19 [70]. 55
Figure II.23 : La couche Inception proposée par GoogleNet pour réduire le nombre de paramètres [73]. 56
Figure II.24 : Classification des méthodes de reconnaissance d'activités humaines. 57
Figure II.25 : Les approches de la fusion pour l'incorporation de l'information temps 58
Figure II.26 : Principe de la méthode proposée dans [14]. 59
Figure II.27 : a) Image de profondeur, b) Front-View BMI, c) Side-View BMI, d) Top-View BMI [14] 60
Figure II.28 : Architecture 3D-ConvNet utilisée pour la construction des descripteurs spatio-temporels [15]. 61
Figure II.29 : Organigramme de la méthode proposée dans [15] 62
Figure II.30 : Organigramme de la méthode proposée dans [16]. 62
Figure II.31 : Exemples d'image généré par rotation des images en profondeurs [16]. 63
Figure II.32 : Exemple de codage pseudo-couleur des WHDMMs[16]. 64
Figure II.33 : Organigramme de la méthode proposée dans [17]. 65
Figure II.34 : Approche proposée dans [13]. 65
Figure II.35 : Descripteurs de chaque canal de la méthode proposée dans [13]. 66
Figure II.36 : Organigramme de la méthode *two stream* [18] 67
Figure II.37 : Exemple du flux optique : a et b) Deux images successives, c) Le flux optique dans la zone en bleu, d) La composante horizontale du flux optique. E) La composante verticale du flux optique [18]. 67

Chapitre III

Figure III.1 : Echantillons de la base de données de Weizmann [5] 72
Figure III.2 : Echantillons de la base de données Keck Gesture dataset [79]. 73
Figure III.3 : Echantillons de la base de données KTH [45]. 74

Figure III.4 : Courbe ROC.....75
Figure III.5 : Interprétation du ROC Curve.....76
Figure III.6 : Matrice de confusion.....77
Figure III.7 : Exemple de la transformée DCT.....78
Figure III.8 : Organigramme de la méthode proposée en utilisant les cartes spatio-temporelles. 79
Figure III.9 : Echantillons d'images et leurs silhouettes tirées de la base de données de Weizmann.....80
Figure III.10 : Cartes spatio-temporelles calculées sur la base de données de Weizmann.....81
Figure III.11 : Organigramme de la méthode proposée en utilisant les silhouettes. 82
Figure III.12 : Matrice de confusion en utilisant 10 activités.....85
Figure III.13 : Matrice de confusion en utilisant 9 activités.....86
Figure III.14 : Projections des vecteurs caractéristiques tirés de la base de données de Weizmann en utilisant PCA. 88
Figure III.15 : Variation du temps de validation en fonction du nombre de neurones.....89
Figure III.16 : Variation du taux de validation en fonction de sigma. 90
Figure III.17 : Courbe d'apprentissage du modèle RBF. 91
Figure III.18 : Matrice de confusion en utilisant une reconnaissance image par image. ..92

Chapitre IV

Figure IV.1 : Organigramme de la méthode proposée.....95
Figure IV.2 : Exemple de détection des personnes dans la base de données de Weizmann : a) Walk, b) Side, c) run. 95
Figure IV.3 : Exemple d'extraction des zones englobantes dans la base de données de Weizmann.....96
Figure IV.4 : Exemple d'extraction des silhouettes à partir des zones englobantes dans la base de données de Weizmann.....96
Figure IV.5 : Exemple de descripteurs BSTM en utilisant la base de données « Kech Gesture Database ». 98
Figure IV.6 : Les 16 activations de la couche de convolution de l'activité « Running » dans la base de données de Weizmann.....99

Figure IV.7 : a) Exemple de BSTM de chaque activités, b) Les descripteurs issus de la couche entièrement connectée (base de données de Weizmann). 100
Figure IV.8 : De gauche à droite, échantillons d'image de la base de données de Weizmann, son BSTM et l'activation de la couche à convolution correspondante. 101
Figure IV.9 : La courbe d'apprentissage en utilisant la base de données de Weizmann. .. 102
Figure IV.10 : Matrice de confusion en utilisant la base de données de Weizmann. 103
Figure IV.11 : Comparaison des descripteurs BSTM de l'ensemble de test des deux activités : a) Run, b) Skip. 103
Figure IV.12 : Courbe ROC en utilisant la base de données de Weizmann. 104
Figure IV.13 : De gauche à droite, échantillon d'image de l'activité Turn left, son BSTM et les activations de la couche de convolution. 104
Figure IV.14 : La courbe d'apprentissage en utilisant la base de données de Keck Gesture Database. 105
Figure IV.15 : Matrice de confusion en utilisant la base de données de Keck Gesture Database. 106
Figure IV.16 : ROC Curve en utilisant la base de données de Keck Gesture Database. ... 106
Figure IV.17 : De gauche à droite : échantillon d'image de la base de données KTH, son BSTM et les activations issues de la couche de convolution. 107
Figure IV.18 : La courbe d'apprentissage en utilisant la base de données KTH. 109
Figure IV.19 : Matrice de confusion en utilisant la base de données KTH. 110
Figure IV.20 : Le ROC Curve en utilisant la base de données KTH. 111

Chapitre V

Figure V.1 : Principe du Système YOLO [83]. 114
Figure V.2 : Exemple de détection d'objets par YOLO [84]. 114
Figure V.3 : Architecture du YOLO [83]. 115
Figure V.4 : Principe de fonctionnement de l'architecture YOLO [83]. 115
Figure V.5 : Architecture détaillée du YOLO V1. 116
Figure V.6 : Organigramme de la méthode proposée. 117
Figure V.7 : Organigramme du protocole de fusion proposé. 118
Figure V.8 : Exemple de reconnaissance en utilisant la base de données de KTH. 120
Figure V.9 : Courbes d'apprentissages en utilisant la base de données KTH.
a) Transfert learning, b) fine-tuning total 121

Figure V.10 : Matrice de confusion en utilisant le fine-tuning total sur la base de données KTH. 122
Figure V.11 : La courbe ROC du modèle YOLO fine-tuné en utilisant la base de données KTH. 123
Figure V.12 : Histogrammes des classes des séquences vidéo de test pour les deux activités : a) Boxing, b) Waving. 124
Figure V.13 : Matrice de confusion en utilisant la base de données KTH. 126
Figure V.14 : Interface graphique de reconnaissance. 127
Figure V.15 : Exemple de simulation en utilisant la base de données de Weizmann et le modèle YOLO fine-tuné. 128
Figure V.16 : Matrice de confusion en utilisant la base de données de Weizmann et le modèle YOLO fine-tuné. 129
Figure V.17 : Matrice de confusion lors de l'utilisation de l'ensemble test de la base de données de Weizmann. 130
Figure V.18 : Résultats de simulation en utilisant des vidéos YouTube. 131
Figure V.19 : Exemple de simulation en temps réel en utilisant la caméra 132

Liste Des Tableaux.

Chapitre III

Tableau III.1 : Protocole de test utilisé dans le cas des descripteurs spatio-temporelles. 83
Tableau III.2 : Taux de reconnaissance lors de l'utilisation des cartes de descripteurs spatio-temporelles. 83
Tableau III.3 : Comparaison des résultats obtenues par rapport aux autres techniques de l'état de l'art. 84
Tableau III.4 : Découpage de la base de données en sous-ensembles pour l'entrainement, la validation et le test 87
Tableau III.5 : Taux de reconnaissance en fonction de la taille de la fenêtre des descripteurs 88
Tableau III.6 : les paramètres optimaux du modèle final. 90

Chapitre IV

Tableau IV.1 : Algorithme de calcul des descripteurs BSTM. 97
Tableau IV.2 : Paramètres de la couche de convolution. 99
Tableau IV.3 : Taux de reconnaissance en utilisant la base de données de Weizmann 101
Tableau IV.4 : Taux de reconnaissance en utilisant la base de données de Keck Gesture Database. 105
Tableau IV.5 : Résultats de reconnaissance en utilisant la base de données de KTH. 108

Chapitre V

Tableau V.1 : Sous-ensembles utilisés dans l'apprentissage. 119
Tableau V.2 : Résultats de reconnaissance en utilisant différents niveaux de fine-tuning. 119
Tableau V.3 : Taux de reconnaissance en utilisant différentes valeurs de seuil T. 123
Tableau V.4 : Taux de reconnaissance en utilisant la base de données de KTH. 125
Tableau V.5 : Taux de reconnaissances en utilisant la base de données de Weizmann 128

Liste Des Abréviations.

SIFT : Scale Invariant Feature Transform

SURF : Speeded Up Robust Features

LBP : Local Binary Patterns

MHI : Motion History Image

MEI : Motion-Energy Images

HOG : Histogram of Oriented Gradients

PCA : Principal Component Analysis

DCT : Discrete Cosine Transform

SVM : Support Vector Machine

RBF : Radial Basis Function

BSTM : Binary Space-Time Map

CNN : Convolutional Neural Network

YOLO : You Only Look Once

COCO : Common Objects in Context

HAR : Human Activity Recognition

HMM : Hidden Markov Models

KNN : K-Nearest Neighbours

MHV : Motion History Volume

STIPs : Space-Time Interest Points

LDA : Latent Dirichlet Allocation

DOG : Difference-of Gaussian

HLAC : Histogram of Local Appearance Context

K-ppv : K-plus Proches Voisins

AI : Artificial Intelligence

MLP : Multi Layer Perceptron

ConvNet : Convolutional Neural Network

ReLU : Rectified Linear units

ILSVRC : ImageNet Large Scale Visual Recognition Challenge

GMM : Gaussian Mixture Mode

RNN : Recurrent Neural Network

LSTM : Long Short-Term Memory

WHDMM : Weighted Hierarchical Depth Motion Maps

DBN : Deep Belief Network

ADI : Average Depth Image

DDI : Depth Difference Image

RBM : Restricted Boltzmann Machine

ROC : Receiver Operating Characteristic

TPR : True Positive Rate

FPR : False Positive Rate

AUC : Area Under the Curve

R-CNN : Region Based Convolutional Neural Networks

AWS : Amazon Web Services

Introduction Générale

« La science consiste à passer d'un étonnement à un autre. »

Aristote
Philosophe

Introduction Générale

Contexte et problématiques

La vision humaine est un système très complexe capable de la captures, l'analyse et de l'interprétation des scènes aperçus par l'œil. Depuis toujours, le rêve des humains est la conception de tels systèmes capables d'interpréter les vidéos capturées par une caméra avec une telle précision, rapidité et efficacité. Depuis que, le domaine de la vision par ordinateur a vu le jour, il n'a cessé d'enregistrer de grands progrès.

Les vidéos constituent une source d'information très riche. C'est le moyen de transfert d'informations le plus utilisé actuellement dans la planète. A titre d'exemple, en 2019, 600 000 heures de vidéos sont téléchargées sur YouTube chaque jour. Plus d'un milliard d'heures de vidéos ont été visualisés. Ces chiffres sont très importants, ce qui confirme que les vidéos gagnent toujours du terrain dans la vie quotidienne des humains.

Cette croissance exponentielle des données vidéo est naturellement accompagnée par des progrès dans les techniques automatisées de traitement et d'exploitation de leurs contenus.

L'un des problèmes majeurs qui reste toujours d'actualité dans la vision par ordinateur est le domaine de la reconnaissance d'activités humaines qui a pour but l'interprétation des activités réalisées par des personnes dans des séquences vidéo. C'est-à-dire doter les machines du pouvoir d'analyse et d'interprétation aient ainsi la prouesse de se rapprocher du système naturel visuel humain. La figure 1 montre cinq images consécutives de quatre activités (Wave, Run, Walk et Bend) réalisées par différentes personnes.

Figure 1 : Exemples d'activités humaines (Cinq images consécutives) tirés de la base de données de Weizmann.

Le domaine de la reconnaissance d'activités humaines est particulièrement difficile vue le nombre de contraintes à surmonter :

- Variations des activités humaines
- Variation des mêmes activités réalisée par plusieurs sujets
- Changement du point de vue de la camera
- Changement du fond
- Variation de la luminosité
- Présence de bruit
- L'énorme quantité de données vidéo

La figure 2, montre un exemple de l'activité « Walk » reposant les différents facteurs variables à surmonter dans le domaine de la reconnaissance d'activités humaines. La même activité est réalisée dans différentes situations : variation de fond, luminosité et de direction, différents facteurs de zoom, changement de vêtement, présence d'ombre.

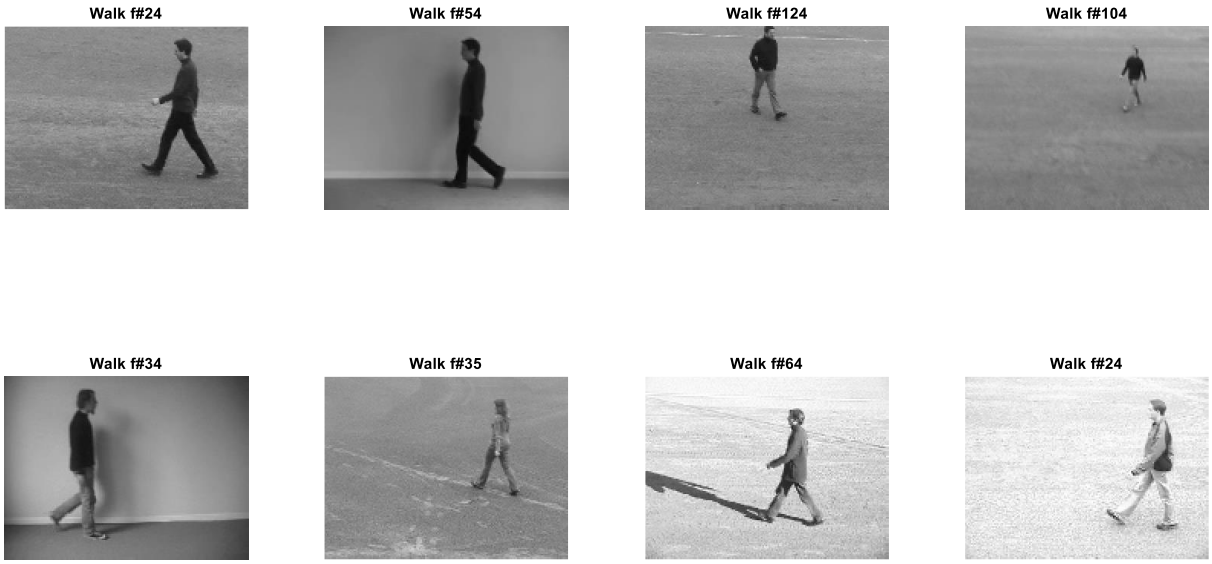


Figure 2 : Exemple de l'activité « Walk » dans différentes situations.

L'importance de cet axe de recherche découle de la grande importance des domaines d'application qui ont bénéficiés des systèmes de reconnaissance d'activités humaines :

La vidéo-surveillance automatisée : ce type de système permet l'analyse automatique des scènes et la détection des activités suspectes telles que l'agression ou l'intrusion. Ce type de système est utilisé dans le domaine des missions militaires de surveillance des infrastructures sensibles, ainsi que dans les drones de surveillance et de reconnaissance. Dans le domaine civil pour la surveillance des personnes âgées dans les maisons de retraites, et la prévention des forces de l'ordre lors d'agression physique sur les personnes physiques et les biens matériels.

L'indexation et la recherche des vidéos : généralement, la recherche des vidéos dans les bases de données à grande échelle tel que le WEB repose sur la description textuelle fait manuellement par les humains. La reconnaissance automatique d'activités humaines permet l'indexation automatique du contenu vidéo ainsi améliorer la pertinence des réponses proposées aux utilisateurs par des moteurs de recherche par exemple.

L'E-santé : la reconnaissance automatique des activités humaines permet la surveillance et l'assistance à distance de malades. Ce type de système permet par exemple de détecter les cas d'urgences tels que la chute et la non prise de médicaments des personnes âgées vivant seules à domicile loin de tous secours immédiats de proximité.

L'interaction homme-machine et la réalité virtuelle : le besoin d'optimisation du temps et de l'argent est la problématique majeure dans le monde de l'industrie. La reconnaissance d'activités humaines a permis la création de système d'interaction homme-machine et réalité virtuelle pour le design de nouveaux prototypes avant de les concrétisés dans la réalité. Ces systèmes sont utilisés dans l'industrie de l'automobile, aéronautique, etc...

Le challenge dans un système de reconnaissance d'activités humaines est l'efficacité à reconnaitre plusieurs activités complexes dans différentes situations avec un taux de reconnaissance élevé, la simplicité du système, la possibilité de l'utiliser en temps réel et l'utilisation de peu de ressources matérielles.

Plusieurs méthodes de reconnaissance d'activités humaines ont été proposées dans la littérature. Au début de cet axe de recherche, les premières techniques étaient basées sur les descripteurs développés dans le cadre de la reconnaissance d'objets, qui ont été adaptées pour le domaine de la reconnaissance des activités humaines. Généralement, ce sont des descripteurs locaux dont la conception est une tâche très délicate. Nous pouvons citer le descripteur de Harris [1], SIFT [2], SURF [3] ou encore le LBP [4]. Ces descripteurs soufrent généralement de plusieurs problèmes tels que la redondance, la difficulté d'extraction et de généralisation.

D'autres techniques de reconnaissance d'activités humaines basées sur les descripteurs globaux ont été aussi proposées. Généralement, ces techniques sont basées sur l'extraction de la forme et la dynamique globale du corps humain, ce sont des descripteurs faciles à extraire, rapides et efficaces. La qualité de descripteur dépend de plusieurs prétraitements tels que la soustraction de fond et la segmentation. Dans ce contexte, nous pouvons citer les travaux de Blank dans [5] qui utilise un volume spatio-temporel, Bobick dans [6] qui a proposé la notion de l'image de l'énergie de mouvement (MEI) et l'image historique du mouvement (MHI), ou encore les techniques basées sur le flux optique. Nous pouvons citer également le descripteur HOG proposé par Dalal et al. dans [7] qui a été combiné avec le PCA par Lu et al. Dans [8] pour créer un descripteur spatio-temporel de reconnaissance d'activités humaines.

Dans d'autres travaux présentés dans l'état de l'art, les auteurs ont essayé de modéliser le corps humain par des représentations simplifiées de la morphologie globale du corps et son évolution dans les différentes activités. Dans [9] Sheikh et al. ont proposé une technique basée sur l'extraction de squelette et son évolution dans un espace XYT. Dans [10], les auteurs projettent les squelettes dans un espace sphérique. Fujiyoshi et al. dans [11] ont proposé la construction d'un squelette simplifie composer de cinq branches. D'autres travaux ont essayé de modéliser le corps humain complètement par un avatar en 3D, dans ce contexte, on trouve par exemple les travaux proposés par Sedai et al dans [12].

Récemment la révolution dans le domaine de l'intelligence artificielle et plus précisément l'apprentissage profond (*Deep Learning*), a permis au domaine de la reconnaissance d'activités humaines d'ouvrir de nouveaux axes de recherche en profitant du pouvoir de tel système à l'extraction des descripteurs automatiquement des données brutes, et d'obtenir des résultats compétitifs par rapport aux techniques conventionnelles de la reconnaissance d'activités humaines. Plusieurs travaux de reconnaissance d'activités humaines basés sur l'apprentissage profond ont été présentés dans la littérature [13-18].

Dans le cadre de cette thèse, nous nous intéressons au problème de la reconnaissance d'activités humaines en utilisant les descripteurs spatio-temporels 2D/3D. Nous allons essayer de présenter cet axe de recherche, son évolution et de proposer des solutions aux contraintes liées à ce domaine.

Contributions

Afin de proposer des solutions aux problématiques citées précédemment dans le domaine de la reconnaissance d'activités humaines, dans cette thèse, quatre différentes approches ont été proposées.

Dans la première approche nous proposons une nouvelle technique de reconnaissance d'activités humaines dans les séquences vidéo, en utilisant les skeletons et la transformée en cosinus discrète (DCT) pour l'extraction des caractéristiques, et la SVM (*Support Vector Machine*) pour la classification des activités.

La deuxième approche est une variante de la première technique, mais dans ce cadre le but et la reconnaissance des activités humaines image par image en temps réel. Cette méthode est basée sur l'extraction des silhouettes et les réseaux de neurones artificiels RBF (*Radial Basis Function*).

Ces deux techniques ont donné de bonnes performances avec des taux de reconnaissance très satisfaisants. Cependant, elles montrent des limitations liées aux algorithmes de classification utilisés. D'un côté la SVM à tendance à donner de faibles performances lorsque le nombre de classes ou la taille de base de données utilisées augmente, de l'autre côté, les réseaux RBF ont tendance à tomber dans les minimas locaux non-optimaux et le sur-apprentissage, de même ils sont limités dans le nombre de neurones dans la couche cachée lors de l'utilisation des problèmes plus complexes.

Pour remédier à ces problèmes, nous proposons une nouvelle technique basée sur l'apprentissage profond (*deep learning*). Nous allons proposer un nouveau descripteur spatio-temporel appelé BSTM (*Binary Space-Time Map*) qui combine l'information temporelle et l'information spatiale de la séquence vidéo. Ces descripteurs BSTM sont utilisés comme entrées au réseau de neurones à convolution CNN (*Convolutional Neural Network*). La méthode proposée a donné de très bons résultats sur toutes les bases de données utilisées. L'étude comparative a montré que ces résultats surpassent ceux obtenus dans les techniques conventionnelles de reconnaissance d'activités humaines et des résultats comparables par rapport aux nouvelles techniques basées sur l'apprentissage profond.

L'étape d'extraction des descripteurs est une opération cruciale dans le domaine de reconnaissance d'activités humaines. Pour surpasser cette limitation, nous proposons une méthode de reconnaissance totalement automatisée. Les opérations d'extraction des descripteurs et de reconnaissance sont réalisées automatiquement par un réseau de neurones à convolution CNN. Dans cette technique, nous proposons l'utilisation de l'apprentissage par transfert (*Transfer learning*) sur le modèle YOLO (*you only look once*) entrainé sur la base de données COCO pour la classification de 80 objets. Cette méthode permet une reconnaissance image par image lors la reconnaissance des activités en temps réel.

Pour adapter cette méthode à la reconnaissance des activités dans les séquences vidéo, nous proposons un protocole de fusion au niveau de l'étape de décision. Le protocole a pour but de délivrer une décision finale unique sur l'activité dans une séquence vidéo à partir des décisions individuelles de chaque image.

Cette méthode a donné des résultats très satisfaisants lors de la reconnaissance image par image, et a surpassé toutes les techniques récentes ou conventionnelles proposées dans la littérature lors de la reconnaissance des activités dans les séquences vidéo.

Plan de la thèse

Le présent manuscrit est divisé en deux grandes parties. Dans la première partie, nous présentons l'état de l'art du domaine. Cette partie est subdivisé en deux chapitres. La deuxième partie est consacré essentiellement à nos contributions. Elle est divisée en trois chapitres.

Partie 1 : Etat de l'art

Chapitre I : Etat de l'Art sur la Reconnaissance d'Activités Humaines

Nous présenterons le domaine de la reconnaissance d'activités humaines, la catégorisation des techniques conventionnelles proposées dans la littérature ainsi que les techniques de classification utilisées.

Chapitre II : Apprentissage Profond Dans La Reconnaissance d'Activités Humaines

Dans le deuxième chapitre, nous détaillons les réseaux de neurones artificiels et de l'apprentissage profond. Ensuite, nous présentons les nouvelles techniques de reconnaissances d'activités humaines basées sur l'apprentissage profond.

Partie 2 : Contributions

Chapitre III : Reconnaissance d'Activités Humaines en utilisant la DCT

Dans le chapitre trois, nous proposons deux techniques de reconnaissance d'activités humaines basées sur la transformée en cosinus discrète DCT. Dans un premier temps, nous présentons une méthode pour la reconnaissance des activités dans les séquences vidéo en utilisant les skeletons et la DCT, ensuite nous présentons une variante de cette technique basée sur les silhouettes et la DCT pour la reconnaissance des activités image par image en temps réel.

Chapitre IV : Reconnaissance d'Activités Humaines en utilisant le Descripteur BSTM et l'apprentissage profond

Nous proposons une nouvelle méthode de reconnaissance d'activités humaines basée sur un nouveau descripteur spatio-temporel appelé BSTM et l'apprentissage profond, nous

présentons l'algorithme d'extraction de notre descripteur ainsi que l'architecture du réseau de neurones à convolution CNN proposée et une étude comparative des résultats.

Chapitre V : Reconnaissance d'Activités Humaines en utilisant le Modèle YOLO

Dans le dernier chapitre, nous proposons une méthode totalement automatisée en utilisant l'apprentissage par transfert (*transfert learning*) du modèle YOLO, dans un premier temps, nous présentons l'architecture du YOLO, ensuite nous détaillerons l'architecture de la méthode proposée pour la reconnaissance des activités image par image en temps réel. Ensuite, nous présentons notre algorithme de fusion pour la reconnaissance des activités dans les séquences vidéo.

Dans ce chapitre, sera présentée également notre interface graphique pour de la simulation de reconnaissance des activités humaines utilisant le modèle YOLO.

Il est à noter que dans chaque chapitre, nous réalisons une étude comparative des résultats obtenus par rapport aux techniques de l'état de l'art.

Nous terminons par une conclusion et quelques perspectives

Partie 1
« Etat de l'Art »

Chapitre I

Etat de l'Art sur la Reconnaissance d'Activités Humaines

« C'est par l'expérience que la science et l'art font leur progrès chez les hommes. »

Aristote
Philosophe

I.1. Introduction

Grâce à l'évolution rapide des systèmes de vision par ordinateur, plusieurs méthodes de reconnaissance d'activités humaines ont été développées ces dernières années. Dans ce chapitre, nous allons présenter le domaine de la reconnaissance d'activités humaines ainsi que l'état de l'art des techniques conventionnelles utilisées. Dans une première étape, nous exposerons les différentes techniques d'extraction de caractéristiques, nous détaillons ensuite les techniques de classification.

I.2. La reconnaissance d'activités humaines

La reconnaissance d'activités humaines (En anglais : *Human Activity Recognition (HAR)*) est un axe de recherche très important et très actif dans le domaine de la vision par ordinateur (*computer vision*), dont le but est de doter les systèmes informatiques de la capacité d'analyser et d'interpréter le mouvement d'un ou plusieurs sujets à partir du contenu visuel d'une scène [19].

La reconnaissance d'activités humaines est un domaine multidisciplinaire qui fait appel à d'autres branches de la vision par ordinateur tel que, la reconnaissance des objets, la segmentation, l'intelligence artificielle et les techniques de classification.

I.3. Terminologie

Plusieurs terminologies ont été utilisées pour la caractérisation d'un mouvement ou la composition de plusieurs mouvements réalisés par un sujet. Moeslund et al. [20] ont proposé une catégorisation selon la complexité des mouvements réalisés :

Le geste : c'est le mouvement élémentaire engendré par le déplacement d'un membre humain tel que, lever la main, tourner la tête, etc.

L'action : c'est un mouvement complexe composé de plusieurs mouvements élémentaires (plusieurs gestes), par exemple courir, tirer un ballon, etc.

L'activité : c'est un mouvement de plus en plus complexe composé de plusieurs actions, par exemple jouer au football, boire du thé, etc.

En général dans la littérature, les deux termes **action** et **activité** désignent la même chose. Dans ce qui suit dans cette thèse, nous avons adopté la même terminologie.

I.4. Le workflow général des techniques de reconnaissance d'activités humaines

Plusieurs techniques de reconnaissance d'activités humaines ont été proposées dans la littérature. Elles sont généralement composées de deux étapes (figure I.1) : 1) Pré-traitement et extraction des caractéristiques, 2) classification des activités.

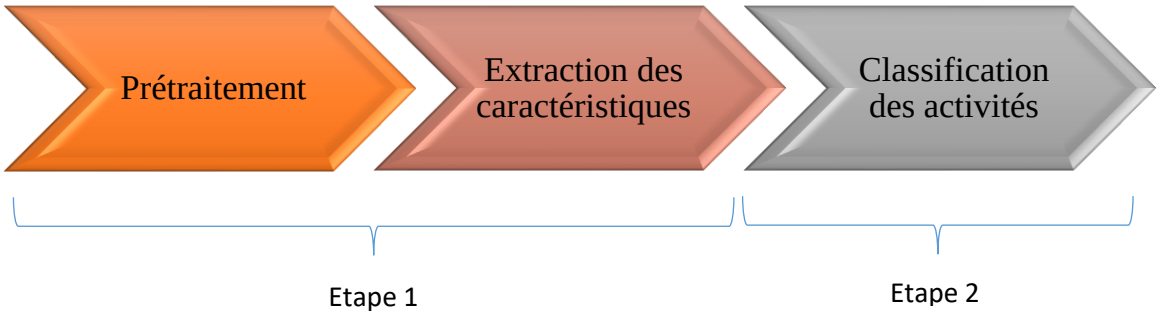


Figure I.1 : Le workflow d'une technique de reconnaissance d'activités humaines.

I.4.1. Pré-traitement et extraction des caractéristiques

Cette étape est composée de deux opérations distinctes :

- **Pré-traitement des données :** le but de cette opération est de préparer les images du flux vidéo pour l'extraction des caractéristiques. Cette étape est composée généralement de différentes opérations de traitement d'images telles que la conversion des images (RGB au niveau du gris, etc.), la normalisation, redimensionnement des images, l'extraction des silhouettes, etc.
- **Extraction des caractéristiques :** c'est l'étape la plus importante dans une technique de reconnaissance d'activités humaines. Le but est l'extraction des vecteurs de caractéristiques séparables qui représentent l'évolution spatio-temporelle du sujet dans la vidéo pour identifier chaque activité d'une façon fiable.

I.4.2. La classification des activités : le but de cette étape est d'associer une étiquette aux activités dans une séquence vidéo représentée par un ensemble de caractéristiques par l'utilisation de classifieur, plusieurs types de classifieurs ont été utilisés dans la littérature. Nous pouvons citer la SVM (*Support Vector Machine*), HMM (*Hidden Markov Model*), réseaux de neurones artificiels, Kalman Filter et les KNN.

I.5. Catégorisation des approches de reconnaissance d'activités humaines

L'une des étapes la plus importante pour la reconnaissance d'activités humaines est l'extraction des primitives qui caractérisent le mouvement dans les séquences vidéo. Nous pouvons distinguer trois types d'approches d'extraction des primitives (caractéristiques) : **les approches basées sur les descripteurs globaux**, **les approches basées sur les descripteurs locaux et celles basées sur la modélisation du corps humain (figure I.2)**.

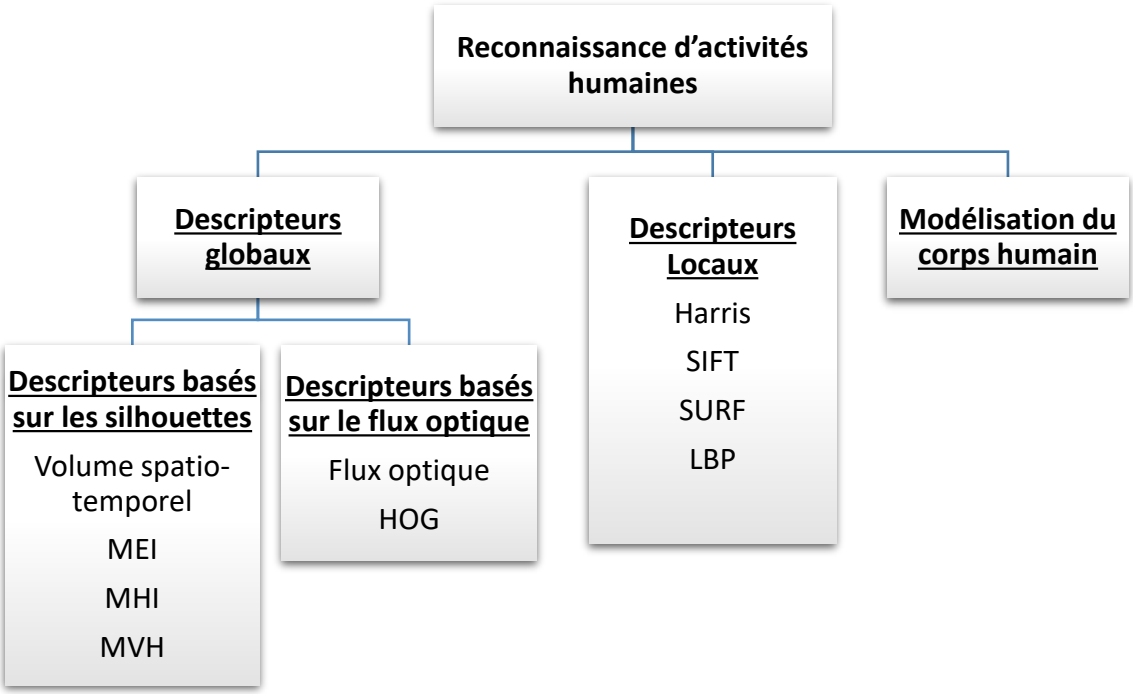


Figure I.2 : Classification des approches de reconnaissance d'activités humaines.

I.5.1. Approches basées sur les descripteurs globaux

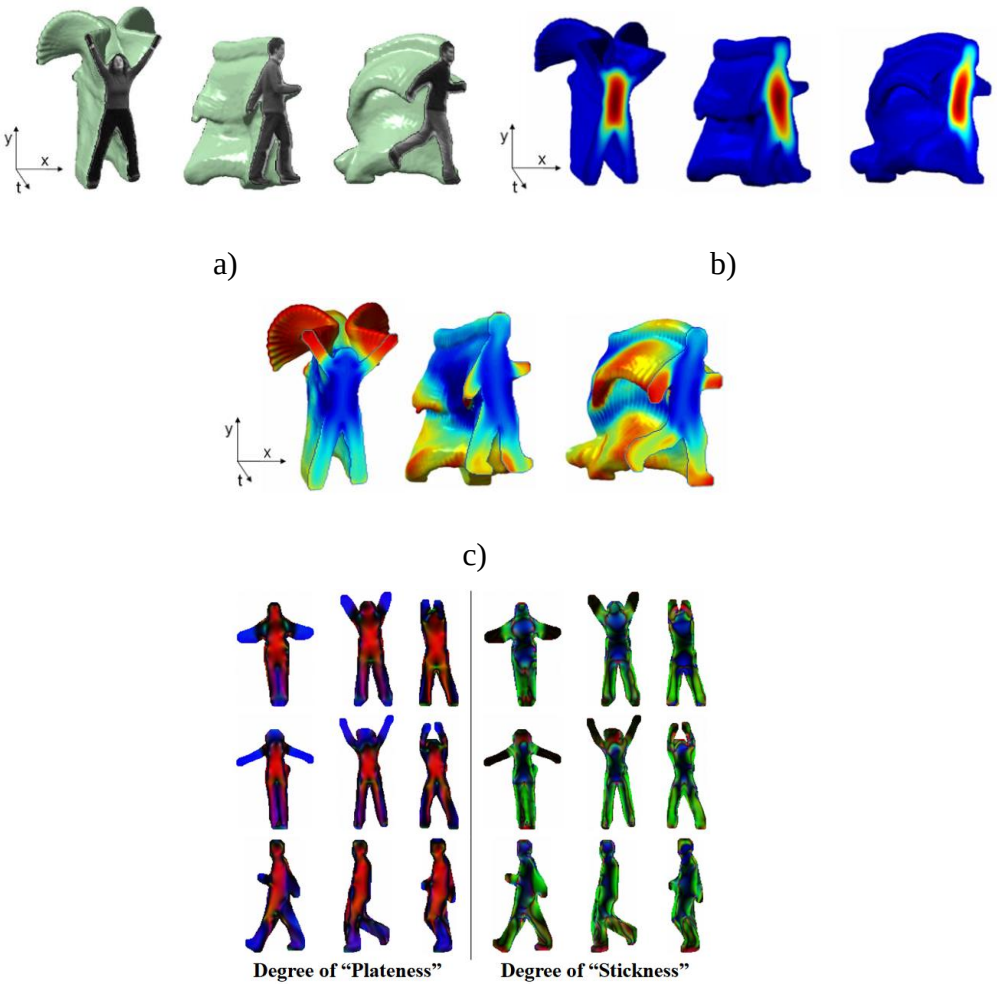
Elles utilisent la structure et la dynamique globale du corps humain. Et modélisent le mouvement et l'apparence globale du corps. Dans ce type de descripteurs, les caractéristiques sont calculées en utilisant l'information sur toute l'image, ils sont généralement simples à extraire, robustes et efficaces [19]. Ces approches sont classées en deux catégories : **approches basées sur les silhouettes** et les **celles basées sur le flux optique et le gradient**.

I.5.1.1. Approches basées sur les silhouettes

L'extraction des caractéristiques en utilisant les approches basées sur les silhouettes nécessite la soustraction du fond. Cette étape est généralement très difficile est la qualité finale du descripteur dépend directement de la qualité des silhouettes soustraites

I.5.1.1.1. Volume spatio-temporel

Les descripteurs à base de silhouettes ont montré leurs efficacités. Dans [5] Blank et al. ont utilisé les silhouettes empilées pour créer un volume spatio-temporel à partir duquel il propose d'extraire les caractéristiques de saillance et d'orientation d'un pixel par rapport à son voisinage en utilisant l'équation de poisson (figure I.3).



a) b)

c)

d)

Figure I.3 : a) Le volume spatio-temporel, b) La solution de l'équation de poisson, c) Les caractéristiques de saillance, d) Les caractéristiques d'orientation [5].

I.5.1.1.2. L'Image d'Energie du Mouvement (MEI) et l'Image Historique du Mouvement (MHI)

Dans [6], Bobick et al. ont introduit le concept du *Template spatio-temporel*. Ils proposent d'extraire les silhouettes dans chaque image de la vidéo, ensuite, ils regroupent les différences entre les images pour construire un volume spatio-temporel. Deux images qui regroupent la variation de l'activité dans l'espace et dans le temps sont créées (figure I.4) : L'image d'énergie du mouvement (MEI : *Motion Energy Image*) et l'image historique du mouvement (MHI : *Motion History Image*).

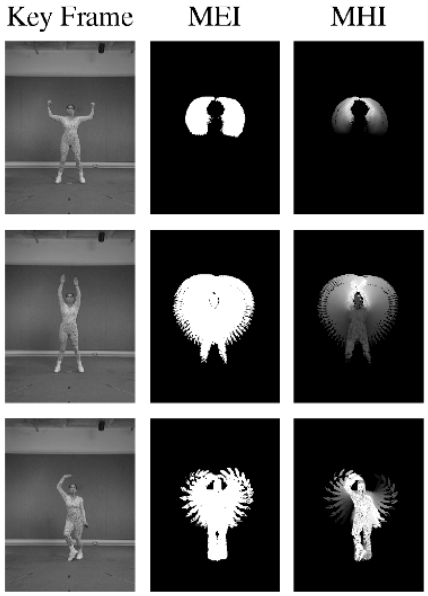


Figure I.4 : Image d'énergie de mouvement (MEI) et l'image de l'historique du mouvement (MHI) [6].

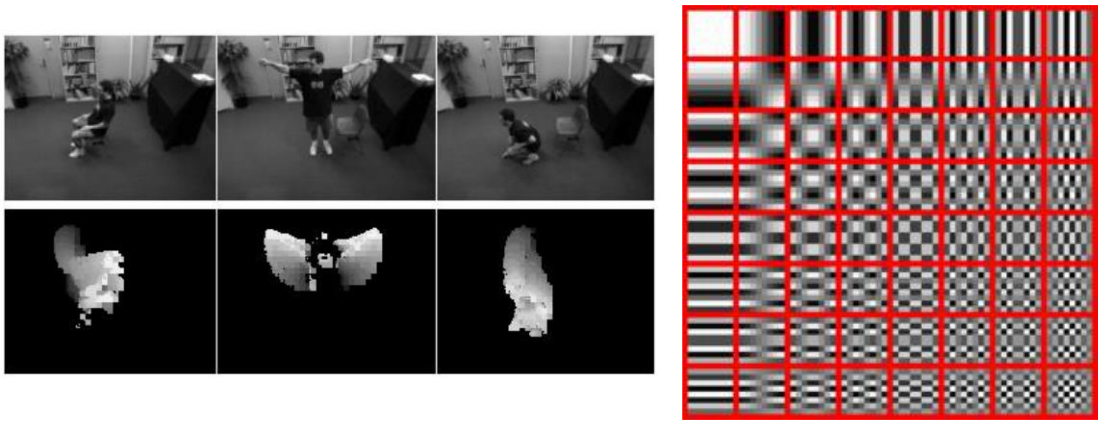
Supposant que $I(x, y, t)$ est une séquence d'image, et $D(x, y, t)$ est la séquence d'images binaires de $I(x, y, t)$, l'image d'énergie de mouvement $E_T(x, y, t)$ est calculée par l'équation suivante :

$$E_T(x, y, t) = \bigcup_{i=0}^{T-1} D(x, y, t - i) \qquad (\text{I. 1})$$

L'image de l'historique du mouvement H_T est calculée par l'équation :

$$H_T(x, y, t) = \begin{cases} T & si\ D(x, y, t) = 1 \\ \max(0, H_T(x, y, t-1) - 1) & sinon \end{cases} \qquad (\text{I. 2})$$

Tasweer et al. ont proposé dans [21] l'extraction des descripteurs en utilisant la transformée en cosinus par block (figure I.5) appliquée sur les images de l'historique du mouvement (MHI).



a) b)

Figure I.5 : a) Image de l'historique du mouvement, b) 2D-DCT par block (8x8) [21].

Dans [22], Weinland et al. ont proposé le volume d'historique du mouvement MVH (*Motion History Volume*) qui représente une extension de l'image de l'historique du mouvement MHI en 3D, le MVH est calculé par la combinaison de plusieurs silhouettes issues de plusieurs caméras.

I.5.1.2. Approches basées sur le flux optique et le gradient

I.5.1.2.1. Flux optique

Contrairement aux approches basées sur l'extraction des silhouettes, les méthodes basées sur le flux optique ne nécessitent pas de soustraction de fonds, les descripteurs sont calculés à partir d'images consécutives.

Le flux optique est défini comme le mouvement apparent de pixels individuels sur le plan de l'image. Il constitue une bonne approximation du véritable mouvement physique projeté sur le plan de l'image. Pour calculer le flux optique, la plupart des méthodes supposent que la couleur/intensité d'un pixel est invariante au déplacement d'une image à l'autre. Le flux optique fournit une description concise des régions en mouvement dans l'image ainsi que la vitesse de celui-ci.

Dans [23] les auteurs ont proposé de diviser le flux optique en quatre champs scalaires différents (figure I.6) : composante positive, négative, horizontale et verticale.

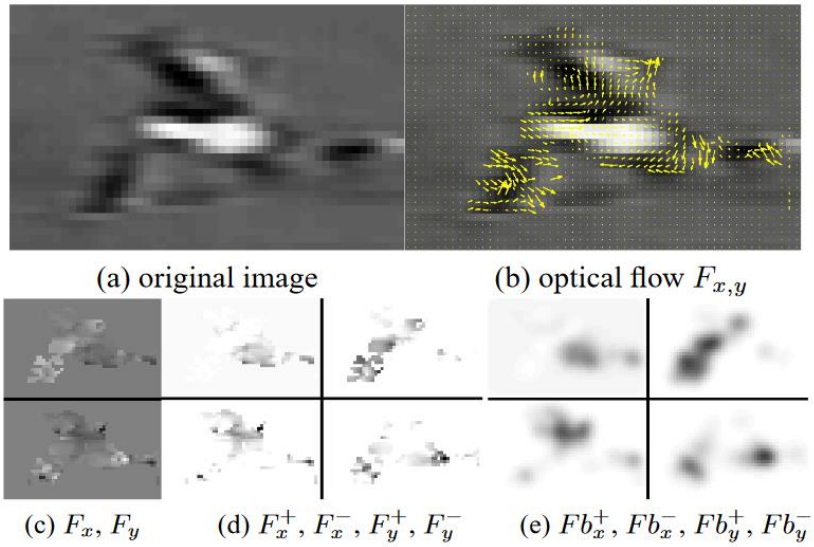


Figure I.6 : a) Image originale, b) Flux optique, c) La séparation de la composante verticale et la composante horizontale, d) Les quatre composantes scalaires, e) Le descripteur final [23].

I.5.1.2.2. Histogrammes de Gradients Orientés (HOG)

L'histogramme de gradients orientés (HOG) est un descripteur qui a été introduit par Dalal et al. Dans [7]. Il définit dans une région les proportions de pixels dont l'orientation du gradient appartient à un certain intervalle. Ces proportions caractérisent la forme présente dans cette région [24].

L'extraction des contours est une étape très importante lors du calcul du gradient, la méthode de Canny a prouvé son efficacité dans ce domaine.

Un filtre gaussien à deux dimensions est tout d'abord appliqué à la luminosité de l'image pour réduire les bruits et limiter le nombre de contours. Ensuite, pour chaque pixel P de l'image, on applique un masque [-1 0 1] puis un masque $[-1\ 0\ 1]^T$. On obtient les valeurs G_P^x et G_P^y. La somme des valeurs absolues des images obtenues en sortie de ces deux filtres ($|G_P^x| + |G_P^y|$) donne une image qui correspond à la magnitude du gradient. En effet, cette valeur représente les encarts de nuances selon les orientations verticales et horizontales (figure I.7).

L'orientation du gradient est obtenue par l'arc tangent du rapport entre la sortie du filtre horizontal et la sortie du filtre vertical $\arctan\left(\frac{G_P^x}{G_P^y}\right)$.

a b c d

Figure I.7 : a) Image initiale, b) Luminance de l'image initial, c) Magnitude du gradient par filtrage de Canny, d) Orientation du gradient par filtrage de Canny (les nuances rouges représentent les orientations verticales et les nuances vertes représentent les orientation horizontales) [24].

La deuxième étape dans le calcul du HOG est le calcul des histogrammes, pour cela, nous considérons comme espace généré par l'orientation du gradient l'intervalle [0, π].
Cet espace est divisé en N_{bin} intervalles. Les intervalles sont de même taille ($\frac{\pi}{N_{bin}}$).
L'ensemble Δ des intervalles d'orientation du gradient est défini par :

$$\Delta = \left\{ I_k = [\frac{2k-1}{2N_{bin}}\pi, \frac{2k+1}{2N_{bin}}\pi], k \in [1, N_{bin}] \right\} \qquad (I.3)$$

Le décompte des pixels de l'orientation du gradient appartenant à un intervalle est binaire. Il est incrémenté de 1 si l'orientation appartient à l'intervalle et de 0 dans le cas contraire. Un pixel dont l'orientation du gradient est proche de la valeur moyenne de l'intervalle caractérise un peu plus celui-ci. Un pixel dont l'orientation du gradient est éloignée de la valeur moyenne de l'intervalle est également caractéristique de l'intervalle voisin. Une interpolation est alors réalisée pour adoucir les contraintes du décompte [24].
Soit D_k, le décompte des pixels pour l'intervalle I_k. sans interpolation, le décompte pour chaque pixel de l'image d'orientation du gradient θ est réalisé par :

$$pour\ tout\ k \in [1, N_{bin}], D_k = \begin{cases} D_k + 1\ si\ \theta \in I_k \\ D_k \qquad\qquad sinon \end{cases} \qquad (I.4)$$

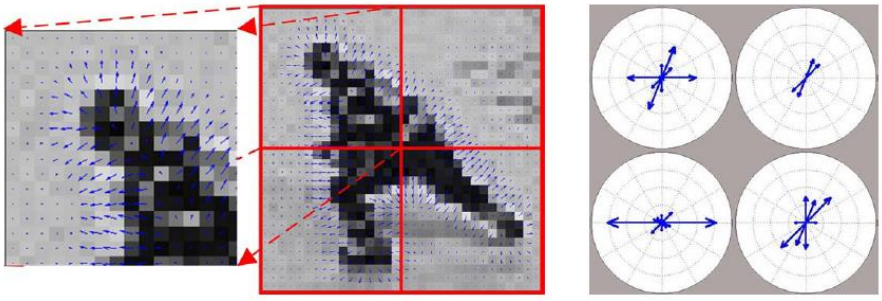
Soit $\theta \in [\frac{l\pi}{N_{bin}}, \frac{(l+1)\pi}{N_{bin}}]$ l'orientation du gradient d'un pixel de l'image. Avec interpolation, la règle de décompte devient :

$$pour\ tout\ k \in [1, N_{bin}], D_k = \begin{cases} D_k + \dfrac{\theta - \dfrac{l\pi}{N_{bin}}}{\dfrac{\pi}{N_{bin}}} & si\ k = l \\ D_k + \dfrac{\dfrac{l\pi}{N_{bin}} - \theta}{\dfrac{\pi}{N_{bin}}} & si\ k = l+1 \\ D_k & sinon \end{cases} \quad (I.5)$$

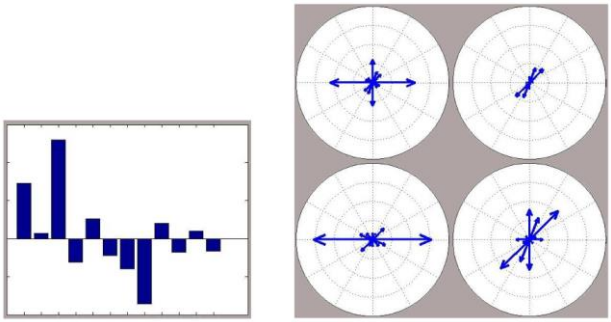
Ces décomptes correspondent aux fréquences d'occurrence des pixels de l'image dont l'orientation du gradient appartient à un intervalle de Δ . Soit v_{I_k}, la fréquence d'apparition relative à l'intervalle I_k, l'histogramme de gradients orientés de la région est alors l'ensemble de ces fréquences pour tous les intervalles de Δ.

$$HOG = \{v_I^{region}, I \in \Delta\} \quad (I.6)$$

Lu et al. Dans [8] ont proposé l'utilisation de l'histogramme de gradients orientés (HOG) initialement utilisait dans la détection de personnes et objets [7], les auteurs proposent de calculer le HOG pour chaque paquet d'images et ensuite d'utiliser le PCA (Principal Component Analysis) pour la réduction de la dimension du descripteur (figure I.8).



a) b)



c) d)

Figure I.8 : a) Gradient de l'image, b) HOG descripteur (grille =2x2) et 8 orientations, c) PCA-HOG descripteur avec 12 Principal Components, d) Le descripteur HOG reconstituer [8].

Les études sur le flux optique et le gradient ont montrées que ce type de descripteurs souffre d'un inconvénient majeur, qui est que dans le calcul. Nous supposons que la différence entre les images consécutives dans la vidéo représente la conséquence du mouvement, alors qu'elles peuvent correspondre au changement dans l'éclairage, du fond et d'autres propriétés dans l'image (figure I.9)

Figure I.9 : L'inconvénient du flux optique, sensibilité au changement du fond.

I.5.2. Approches basées sur les descripteurs Locaux

Comme pour les approches basées sur le flux optique et le gradient, les techniques basées sur les descripteurs locaux ne nécessitent pas de pré-traitement comme la soustraction du fond et l'extraction des silhouettes. Ceci évite la propagation des erreurs durant cette phase. Les primitives de cette catégorie sont caractérisées par l'invariance à l'angle, changement d'apparence des sujets (vêtements) et aux occlusions partielles.

Les premiers travaux en reconnaissance d'actions humaines, à partir de caractéristiques spatiales locales, se sont attachés à utiliser les primitives connues en reconnaissance d'objets (Harris [1], SIFT [2], SURF [3] et LBP [4]). Cette analyse repose sur des informations locales, c'est-à-dire sur un ensemble de pixels présentant une singularité, que ce soit au niveau du gradient ou du contour. Ces points d'intérêts sont donc les caractéristiques spatiales 2D présentent dans les images. Ces primitives sont invariantes à la taille et à l'orientation. Elles fournissent un bon moyen d'identifier des objets dans une image. Les recherches se sont alors orientées vers une utilisation de ces primitives afin d'identifier des comportements humains. Pour cela, les spécialistes du domaine ont appliqué ces descripteurs aux vidéos contenant des actions humaines pour extraire des informations discriminantes sur l'apparence des comportements humains.

I.5.2.1. Descripteur de Harris

Il se base sur une fonction d'autocorrélation du signal c'est-à-dire sur les changements du signal dans plusieurs directions [1]. La figure I.10 montre les différentes étapes du détecteur de Harris. nous commencons par calculer en chaque pixel $p^{i,j}$(i, j) de l'image la matrice d'autocorrélation (ou des moments du second ordre ou tenseur de structure) :

$$M(i,j) = \begin{pmatrix} \mu_{11}(i,j) & \mu_{12}(i,j) \\ \mu_{21}(i,j) & \mu_{22}(i,j) \end{pmatrix} \quad \text{(I. 7)}$$

$$\mu_{11}(i,j) = \sum_{p=-n}^{n} \sum_{q=-n}^{n} w(p,q) I_i^2(i+p, j+q) \quad \text{(I. 8)}$$

$$\mu_{22}(i,j) = \sum_{p=-n}^{n} \sum_{q=-n}^{n} w(p,q) I_j^2(i+p, j+q) \quad \text{(I. 9)}$$

$$\mu_{12}(i,j) = \mu_{21}(i,j) = \sum_{p=-n}^{n} \sum_{q=-n}^{n} w(p,q) I_i^2(i+p, j+q) I_j^2(i+p, j+q) \quad \text{(I. 10)}$$

Où I_i et I_j sont les dérivées premières des niveaux de gris de l'image obtenues par convolution avec les masques de dérivation issus du filtre gaussien et $w(p,q)$ sont des poids de lissage gaussiens tels que $\sum_{p=-n}^{n} \sum_{q=-n}^{n} w(p,q) = 1$.

Nous calculons ensuite les valeurs propres de chaque matrice $M(i,j)$:

$$\lambda_1(i,j) = \frac{1}{2}\left(\mu_{11}(i,j) + \mu_{22}(i,j) + \sqrt{\left(\mu_{11}(i,j) - \mu_{22}(i,j)\right)^2 + 4{\mu_{12}}^2}\right) \quad \text{(I. 11)}$$

$$\lambda_2(i,j) = \frac{1}{2}\left(\mu_{11}(i,j) + \mu_{22}(i,j) - \sqrt{\left(\mu_{11}(i,j) - \mu_{22}(i,j)\right)^2 + 4{\mu_{12}}^2}\right) \quad \text{(I. 12)}$$

Nous pouvons caractériser le pixel $P^{i,j}(i,j)$ de la manière suivante :

- Si les deux valeurs propres sont grandes, alors on est en présence d'un point d'intérêt ;
- Si les deux valeurs propres sont petites, alors le pixel étudié est dans une zone homogène ;
- Si les deux valeurs propres sont très différentes, alors le motif de texture au voisinage du pixel pi,j(i,j) est unidirectionnel.

Le détecteur de Harris calcule pour chaque pixel, la matrice d'autocorrélation à partir des deux composantes des vecteurs gradients de l'image. Ensuite, la matrice de réponse du

détecteur est obtenue à partir de ces matrices. Enfin, les points d'intérêt (figure I.10), sont localisés à partir de cette réponse.

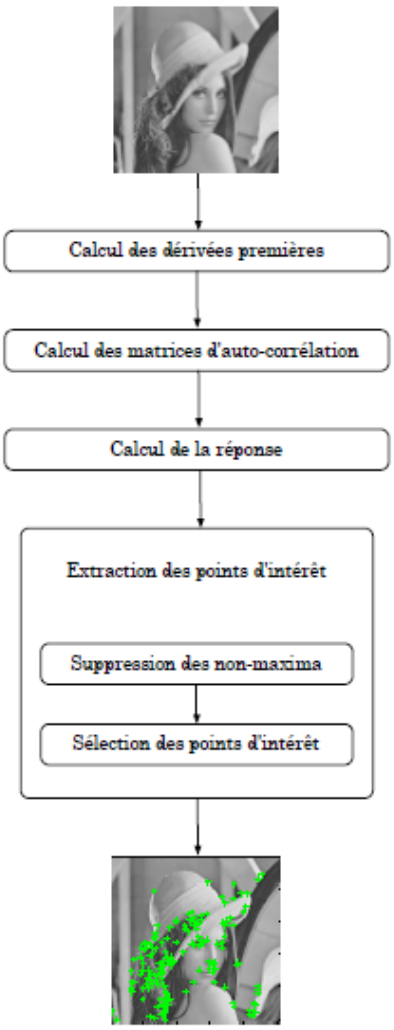


Figure I.10 : Diagramme du détecteur de Harris [1].

Plusieurs travaux dans le domaine de la reconnaissance d'activités humaines en utilisant le descripteur de Harris ont été proposés.

Parmi les premiers travaux proposés pour extraire les points d'intérêts spatio-temporels (STIPs) figure les travaux de Laptev et Lindeberg [25]. Les auteurs ont étendu le détecteur de coins de Harris [1] en lui rajoutant la dimension temporelle. Ce détecteur spatio-temporel, communément appelé Harris 3D, leur permet d'extraire des motifs de mouvement. Ces points d'intérêt spatio-temporel correspondent aux points dont le voisinage local est soumis à une variation spatiale et temporelle significative. L'échelle spatiale et temporelle du voisinage est automatiquement sélectionnée. Ce travail a été amélioré par Laptev et al. [26] pour compenser les mouvements relatifs aux caméras.

Cependant, selon Dollar et al. [27], le nombre des points d'intérêt vérifiant le critère Harris3D est relativement faible par rapport aux zones contenant des mouvements significatifs.

Ainsi, les auteurs ont proposé de nouveaux points d'intérêt spatio-temporels plus denses. Ils appliquent un filtre de Gabor à une dimension au niveau spatial et temporel séparément. Le nombre de points d'intérêt est ajusté en changeant la dimension spatiale et temporelle du voisinage dans lequel les minimas locaux sont sélectionnés. Bregonzio et al. [28] ont étendu cette approche en appliquant un filtre de Gabor 2D au volume spatio-temporel composé par la différence entre les trames adjacentes.

I.5.2.2. Descripteur SIFT (*Scale Invariant Feature Transform*)

Le détecteur SIFT (*Scale Invariant Feature Transform*), présenté dans [2] par Lowe, permet de localiser dans l'image des points clé grâce à un vecteur descripteur, et ce dans le but de caractériser un objet et être capable de le reconnaitre en comparant les caractéristiques des points trouvés à une base de données. L'autre objectif consiste à résoudre le problème du changement d'échelle qui pose généralement des difficultés aux autres détecteurs.

L'algorithme commence par sélectionner des points potentiellement intéressants, invariants aux changements d'échelle et aux rotations, en détectant les extrema locaux, dans l'espace échelle, du Laplacien de l'image, implémenté à l'aide de différences de gaussiennes.

Les points associés à un faible contraste, sont ensuite éliminés en utilisant un seuil sur la valeur du Laplacien après avoir calculé par interpolation la localisation précise des extrema dans l'espace échelle. Parmi les points restants, sont éliminés ceux qui se trouvent sur un contour en fixant un seuil sur le rapport des courbures principales calculé à partir d'une approximation discrète des matrices Hessiennes. Ensuite, une orientation est associée à chaque point d'intérêt en calculant les directions dominantes des vecteurs gradients dans les voisinages des points.

Dans le domaine de la reconnaissance d'activités humaines, Scovanner et al. [29] ont créé une version 3D de l'algorithme SIFT de Lowe pour exploiter le volume spatio-temporel. Dans Ping et al. [30], les auteurs proposent la combinaison du 3D SIFT et le LDA Model (*Latent Dirichlet Allocation*) pour la reconnaissance d'activités humaines.

M. Al Ghamdi et al. ont proposé dans [31] une extension spatio-temporelle de l'algorithme SIFT, en créant une différence pyramidale gaussienne DOG (*Difference-of-Gaussian*) spatio-temporelle pour détecter les maxima locaux (figure I.11).

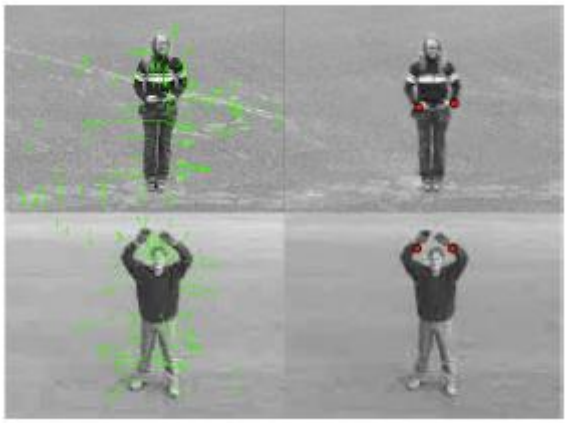


Figure I.11 : Comparaisons entre le 2D SIFT (à gauche) et le ST SIFT (à droite) [31].

I.5.2.3. Descripteur SURF (*Speeded Up Robust Feature*)

Le descripteur SURF est une amélioration du descripteur SIFT. Dans l'algorithme SURF [3], les auteurs utilisent une méthode approchée (par intégrales d'images, introduites dans [32]) pour calculer les dérivées secondes de la convolution de l'image par une gaussienne, et ensuite en déduire le déterminant de la matrice Hessienne. Le descripteur est ensuite calculé. Pour cela, une région rectangulaire alignée sur la direction principale locale est divisée en 16 sous-régions. Dans chaque sous-région, les réponses aux ondelettes de Haar, horizontales et verticales, notées dx et dy, sont calculées en 25 points régulièrement échantillonnés. Pour chacune des 16 sous-régions, le vecteur descripteur est calculé : $v = (\sum dx, \sum |dx|, \sum dy, \sum |dy|)$.

Un vecteur à 64 dimensions est ainsi obtenu. Les auteurs introduisent aussi les SURF-128 ou sont sommés les valeurs positives d'ondelettes de Haar et les valeurs négatives dans deux résultats distincts (on double ainsi le nombre total de valeurs). Ce dernier descripteur est le plus performant dans leurs tests. Les auteurs enregistrent également pour chaque descripteur si le signe du Laplacien (trace de la matrice hessienne) est positif ou négatif (s'il agit d'un blob foncé sur fond clair ou l'inverse). Cela permet ensuite d'accélérer la recherche de correspondances (seule la moitié des descripteurs est testée pour une requête). Cette idée n'existait pas dans l'algorithme initial des SIFT.

Pour la reconnaissance d'activités humaines, Le descripteur « *Speeded Up Robust Features* » (SURF) [3] a été aussi adapté à la représentation 3D par Willems et al. [39] sous le nom de eSURF (SURF étendu). Les cuboïdes 3D sont divisés en une grille de cellules. Chaque cellule est représentée par une somme pondérée de réponses d'ondelettes de Haar uniformément échantillonnées.

I.5.2.4. Descripteur local de motif binaire LBP (*Local Binary Patterns*)

Le descripteur LBP (*Local Binary Pattern*) a été proposé en 1996 par Ojala et al. [4] dans le but de réaliser une classification de textures. Son principe est l'analyse statistique de la texture présente dans une image couleur ou en niveau de gris [34].

Le LBP est calculé sur une région de 3x3 pixels (figure I.12) :

$$LBP(x_c, y_c) = \sum_{n=0}^{7} 2^n s(i_n - i_c) \qquad (I.13)$$

Avec $s(u) = 1\ si\ u > 0\ et\ 0\ sinon, (x_c, y_c)$ les coordonnées du point ou le descripteur est calculé, i_c la valeur de ce point et les i_n les pixels voisins.

Le descripteur LBP est invariant à la variation monotone de la valeur des pixels et par conséquent invariant au changement d'illumination [35].

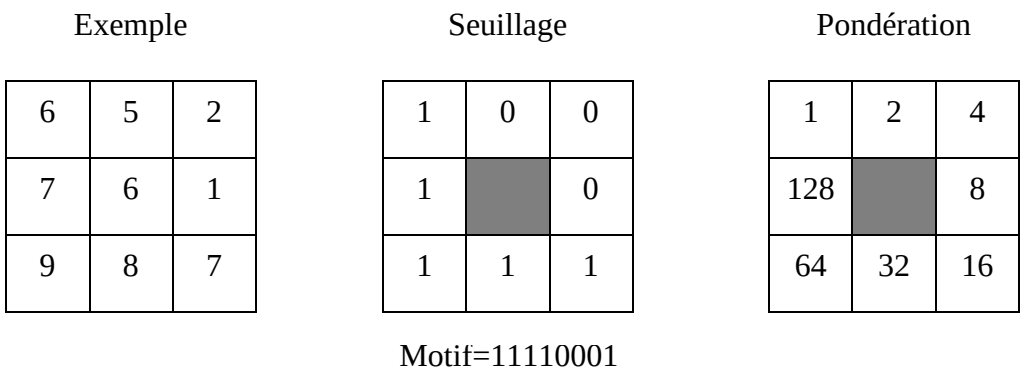


Figure I.12 : Exemple de calcule d'un motif binaire local.

Par la suite l'image est représentée par un histogramme des valeurs LBP de chaque pixel (figure I.13).

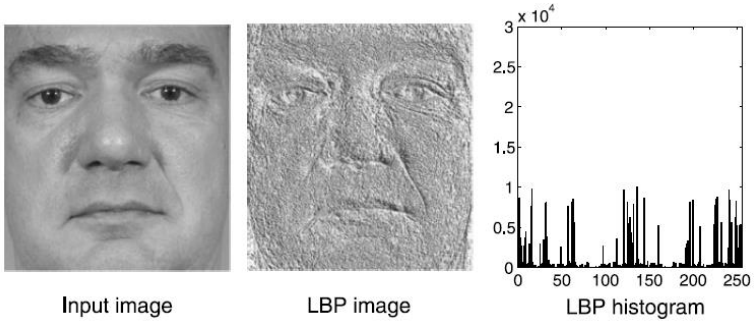


Figure I.13 : Le descripteur LBP [36].

Vili Kellokumpu et al. [37] ont utilisé le descripteur LBP pour la reconnaissance d'activités humaines ; les auteurs ont proposé l'extraction des caractéristiques LBP sur trois niveaux, premièrement au niveau des pixels (chaque échantillon de l'histogramme), ensuite au niveau de chaque région de l'image (histogrammes des sous-volumes), et finalement sur l'image globale après concentration des histogrammes de chaque région. (Figure I.14)

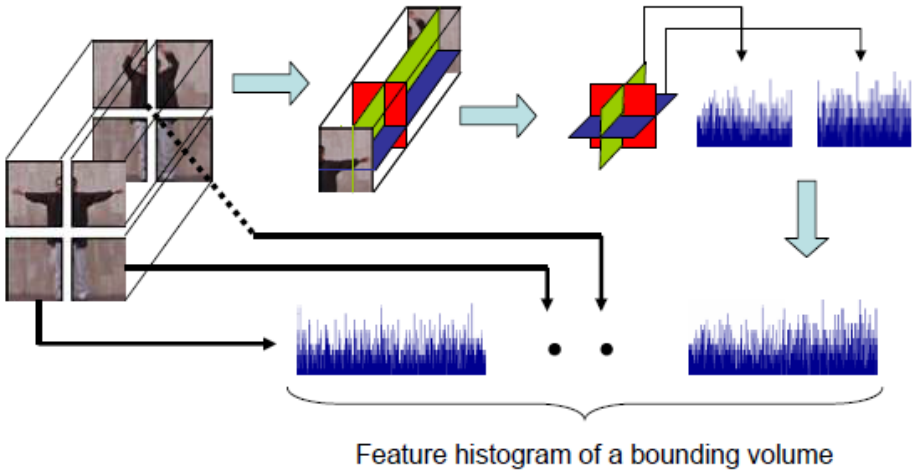


Figure I.14 : La méthode l'extraction des caractéristiques LBP proposée dans [37]

Le descripteur final est la combinaison de tous les histogrammes LBP de chaque portion d'image et pour chaque image de la séquence vidéo (figure I.14).

I.5.3. Les approches basées sur la modélisation du corps humain

Les techniques de reconnaissance d'activités humaines en utilisant la modélisation du corps humain sont basées sur l'étude psychophysique de Johansson réalisée en 1973 [38] et qui a démontré que les humains sont capables d'identifier les activités uniquement à partir de quelques points de repère en mouvement attachés au corps humain (figure I.15).

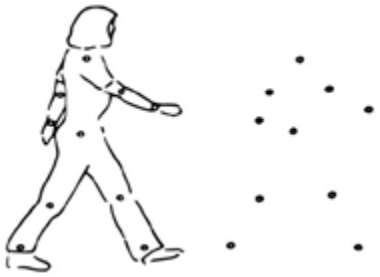


Figure I.15 : Théorie des points de repère de Johansson [38].

Plusieurs techniques de reconnaissance d'activités humaines ont été inspirées de cette théorie, les auteurs ont proposé d'utiliser les différentes parties du corps humain comme point de repère, tel que la tête, les mains, les pieds, etc.

Dans ce cadre, Sheikh et al. [9] ont proposé une technique de reconnaissance d'activités humaines en utilisant les trajectoires des articulations extraites à partir du squelette du sujet (figure I.16).

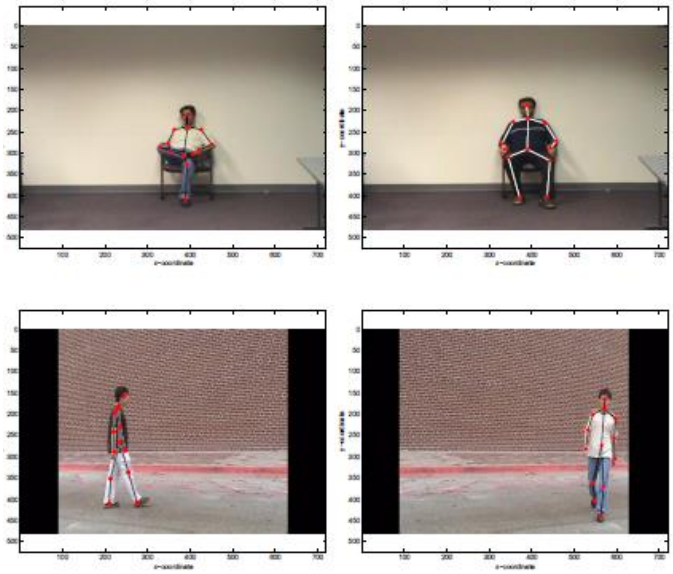


Figure I.16 : Le modèle de squelettes et articulation utilisé par Sheikh et al [9].

Les auteurs proposent la reconstruction de la trajectoire des articulations dans l'espace XYT par la projection sur le modèle de référence de l'activité en utilisant la mesure des angles entre les articulations dans l'espace projeté (figure I.17).

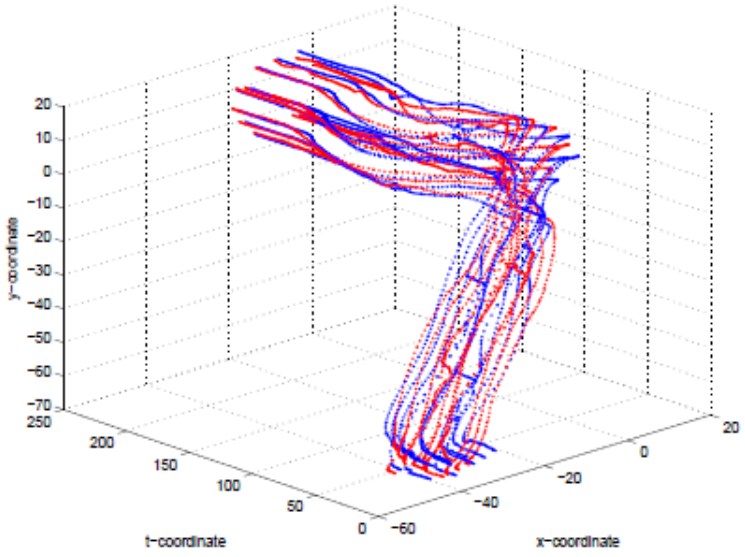


Figure I.17 : La projection de l'activité « Sitting » dans l'espace XYT, les points rouges sont l'action originale, les points Bleus est le modèle reconstruit à partir de la projection sur l'action de base de l'activité « Sitting » [9].

Dans [10], les auteurs introduisent les histogrammes des positions 3D des articulations (HOJ3D), ce type de descripteurs encodent l'occupation spatiale des articulations par rapport au centre de la silhouette, ensuite la position des articulations est projetée dans un espace sphérique échantillonné en plusieurs parties (Bins) et quantifié en utilisant le K-means pour la construction des caractéristiques (figure I.18).

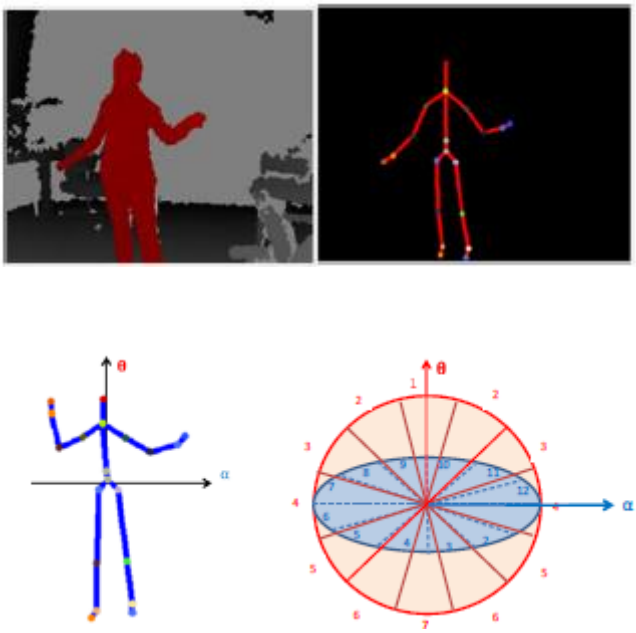


Figure I.18 : a) Image original (profonde), b) squelettes, c) les coordonnées de référence HOJ3D, d) les coordonnées sphériques projetées [10].

Sedai et al. [12] proposent la modélisation du corps humain en 3D pour l'estimation de l'activité humaine (figure I.19) en utilisant le descripteur HLAC (*Histogram of Local appearance Contex*).

Figure I.19 : Le modèle 3D construit par Sedai et al [12].

Fujiyoshi et al. [11] proposent la construction d'un squelette simplifié composé de cinq branches à partir du contour de silhouette de l'objet (figure I.20).

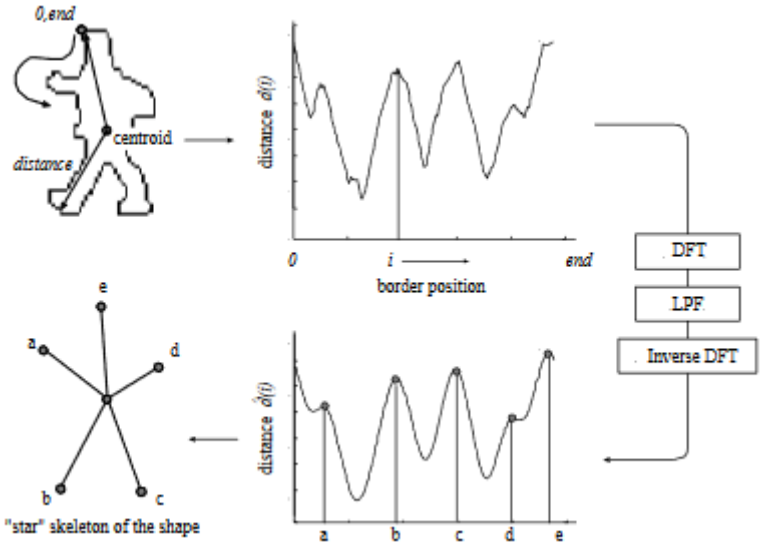


Figure I.20 : Procédure d'extraction de l'Etoile de squelette [11].

L'étoile du squelette (star skeletons) est produite en utilisant les grandes distances par rapport au centroïde du contour de la silhouette.

I.6. Techniques de Classification

Comme mentionné précédemment, les techniques conventionnelles de la reconnaissance d'activités humaines sont divisées en deux étapes, la première est l'extraction des descripteurs caractéristiques des activités et la deuxième étape est l'étiquetage de cette action en utilisant ces descripteurs.

Le but des techniques de classification est de doter la machine de la capacité de simuler le comportement de l'être humain en distinguant les différentes activités par la construction de modèles à partir de données d'apprentissage.

Plusieurs techniques de classification en été utilisées dans la littérature. En général, on peut les classifier en deux familles :

- Méthodes supervisées
- Méthodes non supervisées

I.6.1. Les méthodes de classification supervisées

Les méthodes de classification supervisées sont basées sur la procédure d'apprentissage, l'ensemble d'apprentissage est composée d'un couple entrée-sortie, les entrées sont les vecteurs caractéristiques et la sortie représente l'étiquette correspondant à l'entrée [39].

Le but est donc d'identifier la classe d'appartenance d'une nouvelle entrée qui n'appartient pas à l'ensemble d'apprentissage en utilisant un modèle de classification qui relie l'entrée à la sortie.

I.6.1.1. K plus proches voisins (KNN)

Il s'agit d'un classifieur non-paramétrique où une nouvelle observation est classée dans la classe d'appartenance de l'échantillon d'apprentissage qui lui est la plus proche. L'algorithme KNN est un algorithme qui se base sur le concept de proximité. L'apprentissage par cet algorithme est considéré comme un apprentissage paresseux car il consiste seulement à stocker l'ensemble d'entrainement, cela veut dire, qu'il ne comporte pas d'étape d'apprentissage qui permette d'apprendre un modèle à partir d'un ensemble d'échantillons.

Cet algorithme définit une fonction de distance entre les vecteurs de caractéristiques pour faire la classification d'une nouvelle entrée. Pour classer une nouvelle entrée, l'algorithme procède en identifiant un ensemble de K plus proches voisins pour chaque classe. Cet ensemble

est obtenu en faisant une comparaison entre la nouvelle entrée et chaque exemple de l'ensemble d'entrainement à l'aide d'une mesure de similarité. Une fois que l'ensemble de K plus proches voisins est trouvé, l'algorithme cherche la classe qui a le plus de représentants dans cet ensemble afin d'associer une étiquette à la nouvelle entrée.

La performance du classifieur KNN (ou encore K-ppv) est dépendante de la valeur de K et de la fonction de distance utilisée. L'avantage de cette méthode est qu'elle construit un nouveau modèle pour chaque nouvelle entrée. L'autre avantage est que cette méthode est simple et robuste au bruit. Cette technique devient sensible si le vecteur de grandes dimensions contient un grand nombre de caractéristiques non-pertinentes [39].

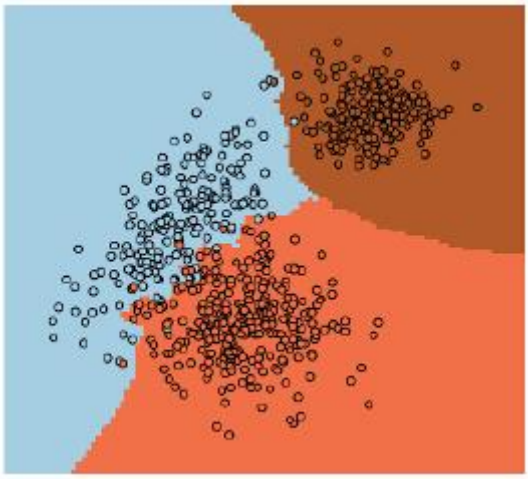


Figure I.21 : Exemple de classification par KNN, classification en 3 classes [40].

La figure I.21 représente une classification de l'espace image à partir de trois classes dans le jeu d'apprentissage. Ici, un vote majoritaire a été utilisé, en tenant compte des 3 voisins les plus proches. Notons que le nombre de voisins à considérer n'est généralement pas corrélé au nombre de classes [40].

Le classifieur KNN et une approche relativement simple, qui a néanmoins donnée de bonnes performances pour la reconnaissance des actions humaines, [21], [5], [41], [42].

I.6.1.2. Machines à vecteurs de support (SVM)

Les Machines à Vecteurs de Support, en anglais *Support Vector Machine* (SVM) élaborée par Vapnik [43] visent à déterminer un hyperplan séparateur entre les espaces des deux classes à séparer. L'idée est de maximiser la marge, c'est-à-dire la distance entre la frontière de séparation et les échantillons les plus proches. Pour cela, l'algorithme transforme l'espace de

représentation des données d'entrée en un espace de plus grandes dimensions, dans lequel il est probable qu'il existe une séparatrice linéaire [24].

L'avantage principal de cette méthode est qu'elle peut être appliquée dans le cas où les classes ne sont pas linéairement séparables. Dans ce cas, les méthodes des SVM tentent de trouver des frontières de décision non linéaire. Le classifieur SVM emploie des fonctions de noyau (polynomial, gaussiennes) permettant de projeter les caractéristiques initiales dans un nouvel espace à grande dimension. Cette projection vise à rendre les données linéairement séparables dans le nouvel espace (figure I.22). Ensuite, le classifieur SVM cherche à trouver un séparateur linéaire dans le nouvel espace qui devient un séparateur non-linéaire dans l'espace original [39].

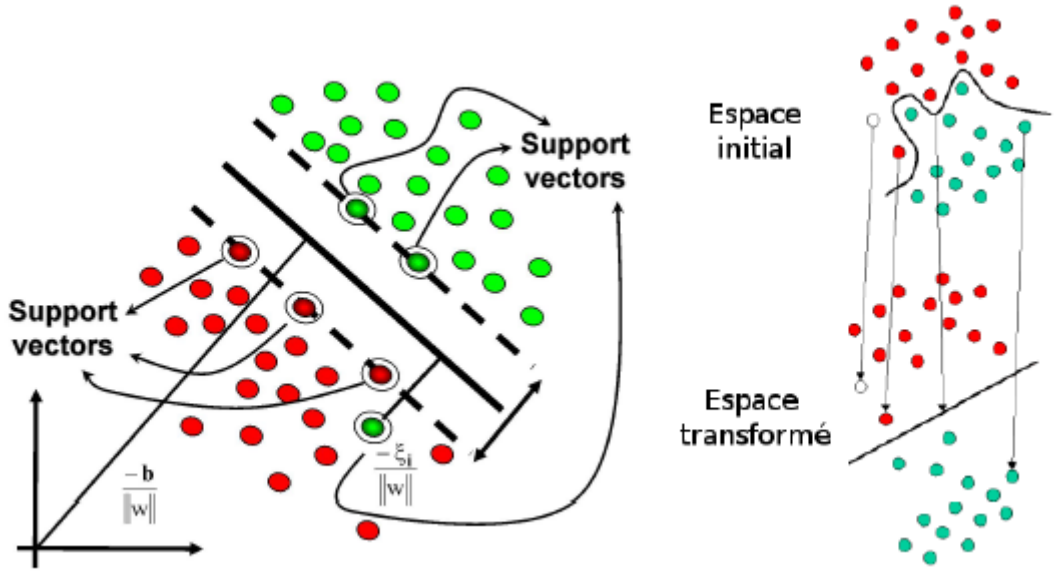


Figure I.22 : a) Cas d'un problème linéairement séparable, b) un cas non linéairement séparable (projection des caractéristiques vers un espace linéairement séparable) [40].

La SVM a été utilisée largement dans le domaine de la reconnaissance d'activités humaines, en donnant de très bons résultats : [44], [45], [46] et [47].

I.6.1.3. Réseaux de neurones

Un réseau de neurones artificiel est un modèle de calcul issu des modèles biologiques. Ce modèle permet de simuler le comportement du cerveau humain. Les réseaux de neurones sont caractérisés par leur capacité d'apprentissage. En général, la structure d'un réseau de neurones est composée d'une succession de couches cachées. La couche d'entrée est reliée à la

couche de sortie du réseau à travers les couches cachées selon une architecture définie, comme l'architecture du perceptron multicouche.

Chaque couche cachée est constituée d'un certain nombre de neurones reliés à la couche précédente. La couche suivante reçoit en entrée les sorties de la couche précédente du réseau.

Le vecteur d'entrée pour chaque couche est pondéré par un poids. À l'aide d'une fonction d'activation, le réseau procède au calcul de ces poids afin de produire une sortie (figure I.23) [39].

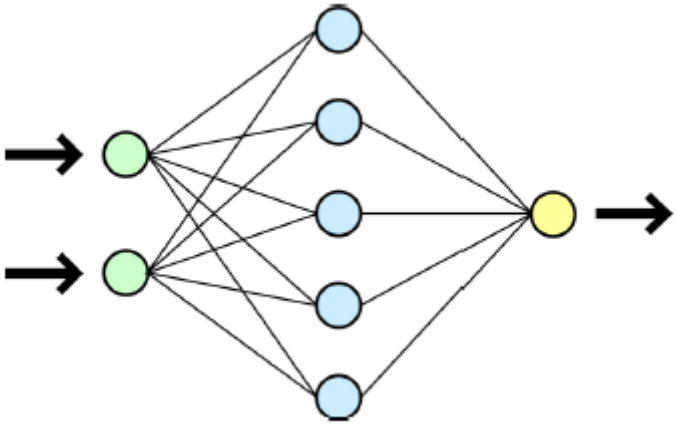


Figure I.23 : Exemple d'un réseau de neurones.

L'objectif dans un réseau de neurones artificiels supervisé est d'entrainer un modèle de neurones qui cherche à s'optimiser vers une sortie précise à partir d'un vecteur de caractéristiques d'apprentissage associé à un vecteur objectif.

L'apprentissage consiste à mettre à jour les poids du réseau avec la comparaison entre la sortie générée et la sortie objective, le réseau s'adapte jusqu'à ce que la sortie corresponde à la cible.

La classification par les réseaux de neurones supervisés a été mise en œuvre dans plusieurs travaux de reconnaissance d'activités humaines : [48] et [49].

I.6.2. Les techniques de classification non supervisées

Les méthodes de classification non supervisée sont des méthodes qui cherchent à identifier, ou à partitionner un ensemble de données à un certain nombre de classes distinctes à partir d'un fichier de description, tout en essayant d'optimiser un critère qui vise à regrouper les données les plus homogènes dans chaque classe.

I.6.2.1. K-moyennes (k-means)

C'est une méthode itérative qui a été proposée par MacQueen en 1967, qui prend comme représentant de chaque classe son centre de gravité, les centres sont mis à jour à chaque nouvelle affectation d'un élément à une classe. L'objectif du k-means est de diviser l'espace des échantillons en k classes (k connu) par amélioration à chaque itération en minimisant la variance intra-classe [50].

Le principe de l'algorithme K-means est le suivant [51]

1. Initialisation
   a) On a les données $X_{1:N}$
   b) Choisir arbitrairement un cluster initial $m_{1:k}$
2. Répéter
   a) Assigner chaque point de données au cluster le plus proche

$$Z_n = argmin_{i \in 1 \ldots k} d(X_n, m_i) \qquad (I.14)$$

   b) Calculer la distance moyenne entre toutes les coordonnées attribuées au nouveau cluster et la moyenne du cluster

$$m_k = \frac{1}{N_k} \sum_{n:Z_n=k} X_n \qquad (I.15)$$

3. Jusqu'à ce que l'affectation de $Z_{1:N}$ ne change pas.

La fonction objective de l'algorithme K-means est la distance entre la somme carrée de chaque point et sa moyenne respective :

$$F(z_{1:N}, m_{1:k}) = \frac{1}{2} \sum_{n=0}^{N} \|X_n - m_{Zn}\|^2 \qquad (I.16)$$

Le point faible de l'algorithme K-means est le choix des conditions initiales qui peuvent affecter les résultats de classement, c'est-à-dire, que la convergence de l'algorithme dépend directement de centres initiaux et du nombre de clusters.

I.6.2.2. Les Modèles de Markov à états cachés (Hidden Markov Models (HMM))

Les modèles de Markov sont une des solutions les plus naturelles pour traiter des données séquentielles. Nous cherchons à mettre en avant des motifs dans la séquence temporelle observée, en corrélant des observations proches de celle-ci. L'exemple commun est celui du temps [40].

Les HMM *(Hidden Markov Models)* font partie des méthodes de classification les plus utilisées dans le domaine de la reconnaissance d'activités humaines, ce sont des modèles génératifs dont la mesure se fait par l'intermédiaire d'états cachés, c'est-à-dire, on ne mesure pas l'état actuel dans lequel le modèle se trouve, mais un état caché qui renseigne sur l'état courant dans lequel le système doit être. Deux mesures sont faites (figure I.24) :

a) Le calcul de la probabilité de transition entre deux états dans le modèle de Markov.
b) Le calcul de la probabilité d'observation d'un état par l'intermédiaire de l'état caché.

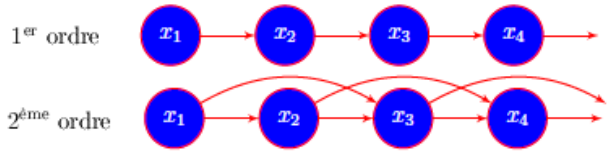


Figure I.24 : Exemple d'une chaine de Markov [40].

Pour rendre le système applicable dans la pratique, deux hypothèses sont faites :

a) L'hypothèse de Markov : la dépendance d'un état dans le temps ne dépend que de l'état précédent.
b) Les observations ne dépendent que de l'état courant, c'est-à-dire chaque observation est indépendante.

Les modèles de Markov à états cachés ont été utilisés dans de nombreuse techniques de reconnaissance d'activités humaines pour modéliser la dépendance temporelle. Yamato et al. Dans [52] ont utilisé les HMM pour classifier les actions de tennis, Weinland et al. dans [53] ont construit un dictionnaire des actions en utilisant les modèles de Markov. D'autres auteurs ont proposé d'utiliser un HMM pour chaque partie du corps tell que dans [54] et [55].

I.7. Synthèse

Les techniques conventionnelles de reconnaissance d'activités humaines peuvent être classées en trois catégories selon le type de descripteur utilisé pour la caractérisation des activités.

Les approches basées sur les descripteurs globaux utilisent la forme et la dynamique globale du corps humain. Ces descripteurs sont faciles à extraire, robustes et efficaces. Ces techniques sont classées en deux catégories : les approches basées sur les silhouettes et les approches basées sur le flux optique.

Les travaux de l'état de l'art ont montré que la qualité des descripteurs basés sur les silhouettes dépend de l'opération de soustraction de fond. Cette dernière est tributaire de plusieurs facteurs tels que, la qualité des images, leurs complexités et l'efficacité d'algorithme d'extraction utilisé.

Contrairement aux techniques basées sur l'extraction des silhouettes, les méthodes basées sur le flux optique ne nécessitent pas un pré-traitement de soustraction de fond, les descripteurs sont calculés directement à partir des images consécutives. Ce type de descripteur a été utilisé avec succès dans plusieurs techniques de reconnaissance. Mais le problème majeur avec le flux optique est la sensibilité au changement du fond.

La deuxième catégorie des techniques de reconnaissance d'activités humaines est basée sur les descripteurs locaux, contrairement aux techniques précédentes, ce type de méthode ne nécessite pas un prétraitement ce qui évite la propagation d'erreurs. Généralement, l'étape d'extraction de ces descripteurs repose sur les concepts développés dans le domaine de la reconnaissance d'objets.

Les descripteurs locaux sont invariants à la rotation, occlusions partielles et l'apparence des sujets. Ils permettent une très bonne représentation des objets dans une image.

Le problème avec ce type de descripteurs, c'est qu'ils sont conçus préalablement pour la reconnaissance d'objets dans les images et ont été adaptés pour la reconnaissance des activités, par la caractérisation de leur variation dans des images consécutives.

Les approches basées sur la modélisation du corps humain représentent la troisième catégorie des techniques de reconnaissance d'activités humaines. Ces types de méthodes sont basées sur la modélisation du corps humain et son évolution dans le temps. Généralement, l'opération de conception du modèle est très complexe et la qualité des descripteurs finaux

dépend de la complexité du modèle du corps humain généré. Ce type de technique nécessite d'importantes ressources de calcul et des vidéos de haute résolution pour une meilleure construction des modèles.

Dans cette thèse, nous proposons deux techniques basées sur les descripteurs globaux, dans la première méthode nous proposons l'utilisation de la transformée en cosinus discrète DCT pour l'extraction des caractéristiques à partir des silhouettes extraites des séquences vidéo. Contrairement aux techniques proposées dans [5] et [6] qui sont basées sur le calcul de plusieurs types de descripteurs combinés, notre méthode est plus rapide, plus simple et adaptée pour la reconnaissance des activités en temps réel.

Dans la deuxième méthode, nous nous sommes inspirés du descripteur MHI (*Image d'historique du mouvement*) présenté dans [6] pour la conception d'un nouveau descripteur appelé BSTM (*Binary Space-Time Map*), ce dernier est une image binaire calculée sur les silhouettes segmentées et centrées sur le sujet. Contrairement aux descripteurs MHI, le descripteur BSTM combine l'information spatio-temporelle de l'activité dans un intervalle de temps des silhouettes de chaque sujet séparément et non pas sur l'image globale, notre méthode est rapide, efficace et applicable pour la reconnaissance de plusieurs sujets dans la même séquence vidéo.

I.8. Conclusion

La reconnaissance d'activités humaines est un domaine de recherche très actif qui a bénéficié de l'évolution des techniques d'extraction des caractéristiques ainsi que des nouvelles techniques de classification.

L'un des outils les plus intéressants et plus importants actuellement dans le domaine de la vision par ordinateur est l'intelligence artificielle et le deep learning. Cet outil très puisant à ouvert de nouveaux horizons dans le domaine de la reconnaissance d'activités humaines.

Dans le chapitre suivant, nous allons présenter une étude détaillée sur les réseaux de neurones artificiels et l'apprentissage profond (*deep learning*).

Partie 1
« Etat de l'Art »

Chapitre II

Apprentissage Profond Dans La Reconnaissance d'Activités Humaines

« Le jour où la science commencera à étudier les phénomènes non physiques, elle fera plus de progrès en une décennie que dans tous les siècles précédents de son existence. »

Nikola Tesla
Ingénieur, Inventeur, Physicien, Scientifique (1856 - 1943)

II.1. Introduction

L'avènement de l'apprentissage profond ou « *Deep Learning* » a révolutionné le domaine de l'apprentissage artificiel. Il a particulièrement permis d'améliorer la capacité d'apprentissage de la machine grâce à l'utilisation de données massives et hétérogènes. L'apprentissage profond découle directement des réseaux de neurones artificiels multicouches. Nous commencerons donc ce chapitre par donner quelques notions de base sur ces derniers. Ensuite, nous présenterons les nouvelles méthodes de reconnaissance d'activités humaines basées sur l'apprentissage profond.

II.2. Réseaux de neurones artificiels

Les réseaux de neurones artificiels sont un outil incontournable dans le domaine de la vision par ordinateur. Ils ont été développés dans les années 40 par analogie avec le système nerveux humain, Ils constituent l'un des éléments de base de l'intelligence artificielle (AI).

II.2.1. Le perceptron (Neurone formel)

L'élément de base dans un réseau de neurones artificiels est le perceptron qui est directement inspiré des neurones biologiques (figure II.1), il est caractérisé par :

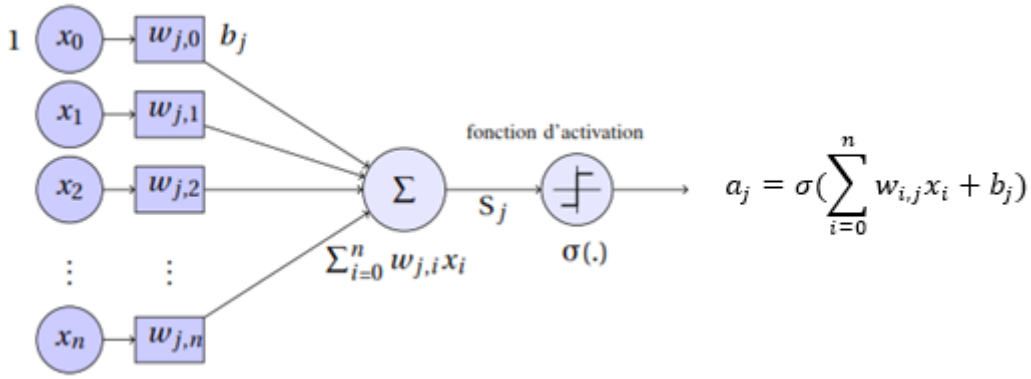


Figure II.1 : Schéma de fonctionnement d'un perceptron [56].

- Son activation (état) a_j : C'est la valeur de sortie du neurone calculé par l'équation ci-dessous [56] :

$$a_j = \sigma(\sum_{i=0}^{n} w_{i,j}x_i + b_j) \qquad \text{(II. 1)}$$

a_j : Est calculée pour un neurone j et un vecteur d'entrée de taille n, elle effectue la somme pondérée par n poids $w_{i,j} \in [0:n]$, ajoute un biais b_j et applique une fonction d'activation σ.

- Ses connexions d'entrée associées à des poids wij (j est l'indice du neurone partageant la connexion)
- Sa fonction d'entrée généralement une somme pondérée.
- Sa fonction d'activation (fonction de transfert) qui calcule à partir de la valeur de la fonction d'entrée l'activation du neurone, elle peut être (figure II.2) : une fonction binaire à seuil, une fonction linéaire à seuil ou bien une fonction sigmoïde [57].

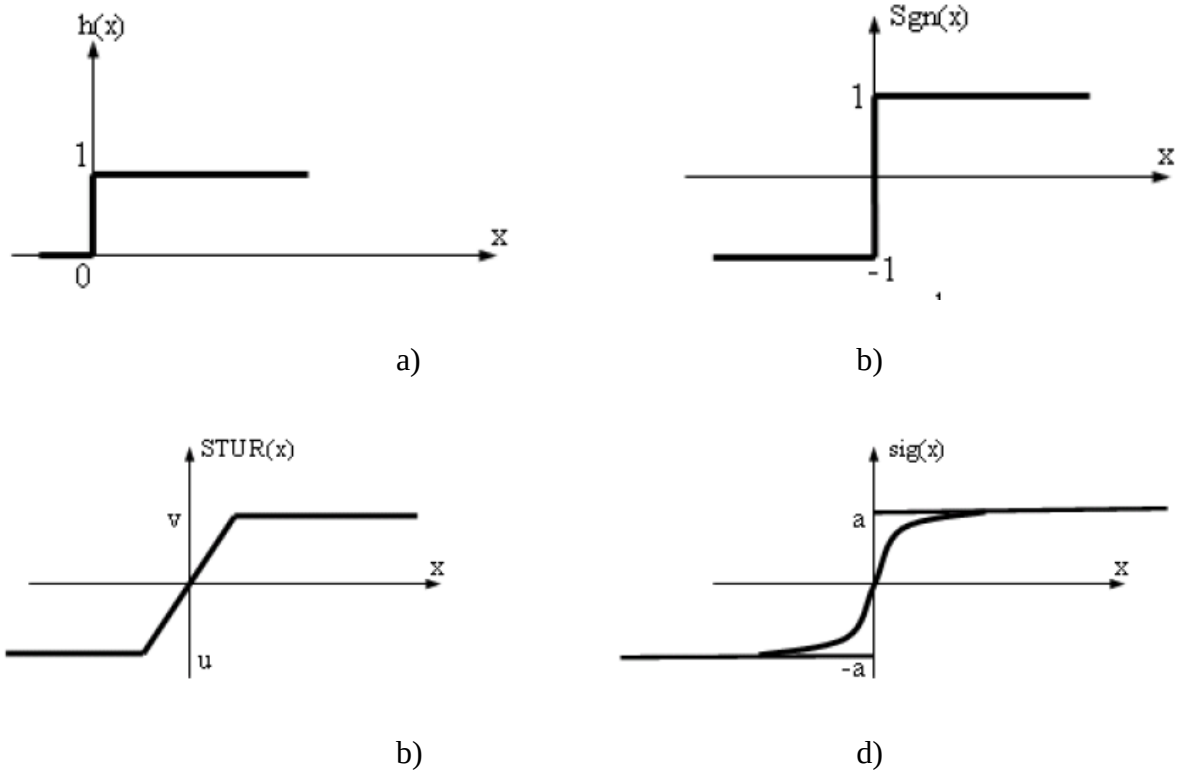


Figure II.2 : Fonctions d'activation [58] : a) Heaviside, b) signe, c) Linéaire à seuil, d) Sigmoïde.

Un réseau de neurones artificiel est l'ensemble de neurones élémentaires (perceptron) reliés entre eux par des connexions pondérées, il se présente sous différentes topologies et se caractérise par [58] :

- La couche d'entrée (neurones d'entrée) : reçois les données d'entrée.
- Les couches cachées : représentent la fonction de transfert du réseau.
- La couche de sortie (neurones de sortie) : fournie les résultats du traitement.

La capacité d'apprentissage est une caractéristique distinctive de certains types de réseaux de neurones artificiels, elle consiste à minimiser l'erreur de prédiction par la mise à jour des poids des connexions pour adapter le résultat selon le problème traité. Deux familles d'apprentissage se distinguent :

Apprentissage supervisé : nécessite de disposer d'un ensemble de couples de données (entrées, sorties désirées correspondantes), la différence entre la sortie du réseau et la sortie désirée donne une erreur utilisée pour réaliser l'adaptation des poids jusqu'à l'optimisation du modèle [57].

Apprentissage non supervisé : aucune sortie n'est disponible, l'ajustement des paramètres dépend des critères internes du réseau et des propriétés communes entre les données d'entrée. Ainsi, après la phase d'apprentissage, lorsqu'une entrée est présentée au réseau, la sortie indique son appartenance à une classe [57].

II.2.2. Perceptron Multicouche (*Multi Layer Perceptron* : MLP)

Pour surmonter les limites du perceptron simple à résoudre les problèmes non-linéaires, un nouveau type de réseaux de neurones artificiels a été proposé dans les années 80. Son principe consiste à superposer plusieurs perceptrons ayant comme fonction d'activation la fonction sigmoïde (figure II.3) [59].

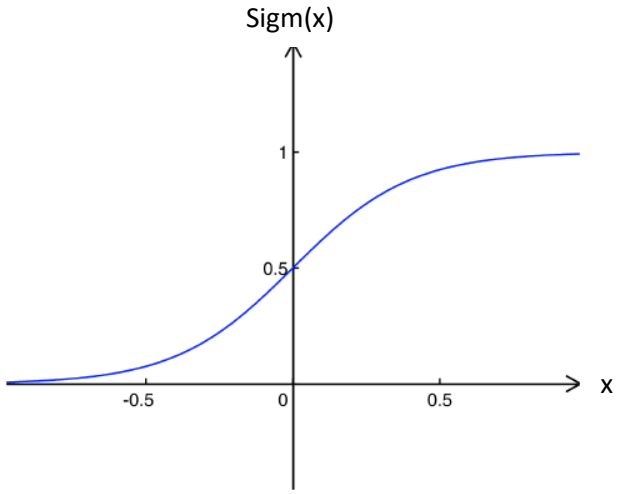


Figure II.3 : Fonction d'activation dans un réseau MLP [59].

$$sigm(x) = \frac{1}{1 + e^{-x}} \qquad \text{(II. 2)}$$

Un perceptron multicouche est composé d'un ensemble d'unités de traitement, connectées entre elles par des connexions pondérées (figure II.4). Les poids de ces connexions sont les paramètres du modèle a optimisé. La figure II.4 montres l'exemple d'un réseau de neurones MLP composé d'une première couche appelée couche d'entrée, deux couches cachées et une couche de sortie.

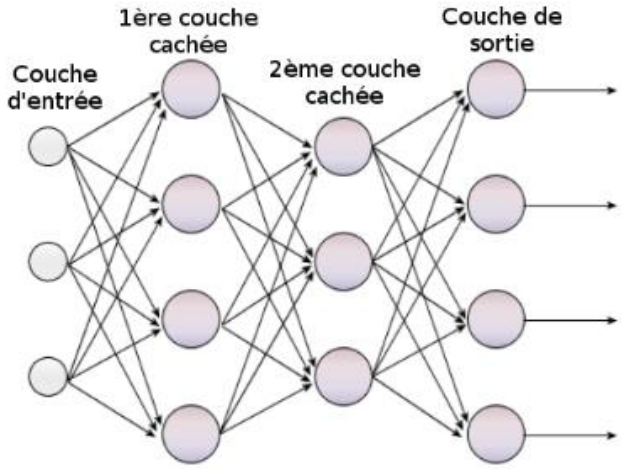


Figure II.4 : Schéma d'un perceptron Multicouche [56].

L'apprentissage du réseau MLP se fait par un algorithme de back-propagation (propagation du gradient) qui représente une descente de gradient sur la fonction d'erreur quadratique entre la sortie et la sortie objective par un apprentissage supervisé, d'où la nécessité de connaître pour chaque vecteur d'entrée, un vecteur de sortie désiré. Cet algorithme opère en deux étapes :

Etape de propagation du vecteur d'entrée : cette opération consiste à calculer le vecteur de sortie du réseau en fonction du vecteur caractéristique d'entrée [59].

Etape de rétropropagation du gradient : cette étape permet le calcul du gradient de la fonction d'erreur quadratique par rapport à chaque paramètre du réseau puis la mise à jour de ces paramètres [59].

L'algorithme de rétropropagation est représenté par la figure II.5 suivante :

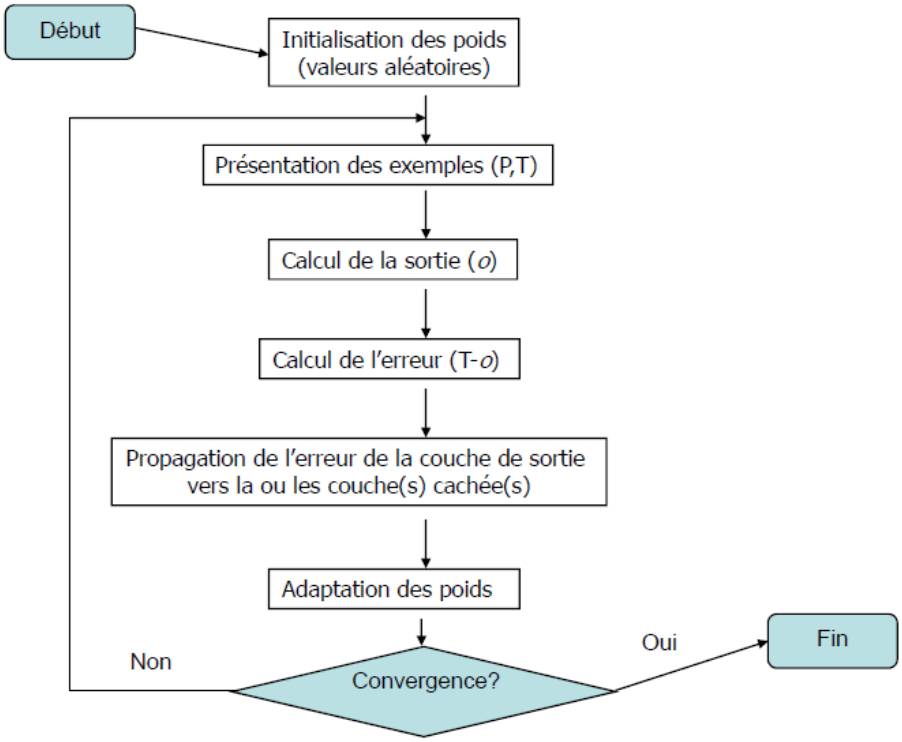


Figure II.5 : Algorithme de rétropropagation [59].

Les réseaux de neurones à perceptron multicouche ont permis de résoudre plusieurs défis notamment la classification des problèmes non linéaires. Cependant le MLP souffre d'un problème major, qui est que si l'architecture est trop profonde, alors l'optimisation des paramètres mène bien souvent à des minimas locaux non optimaux [59], de plus le nombre de paramètres d'apprentissage augmentent ce qui rend l'apprentissage très lent (nombre de couches cachées très limitées).

Pour contourner et remédier aux limitations des réseaux de neurones artificiels, le *deep learning* a été développé, permettant de créer des modèles plus complexes avec plusieurs couches cachées et des millions de paramètres à optimiser.

II.3. L'apprentissage profond (*Deep Learning*)

Dans les techniques d'apprentissage classiques, le but est de trouver les caractéristiques représentatives qui contiennent l'information utile pour la classification, ainsi pour chaque problème étudié, plusieurs types de descripteurs peuvent être combinés pour trouver les bons éléments. Le problème est que généralement ces descripteurs sont incomplets, redondants et très complexes à trouver. La figure II.6 montre la différence entre le *deep learning* est les méthodes conventionnelles du *machine learning*.

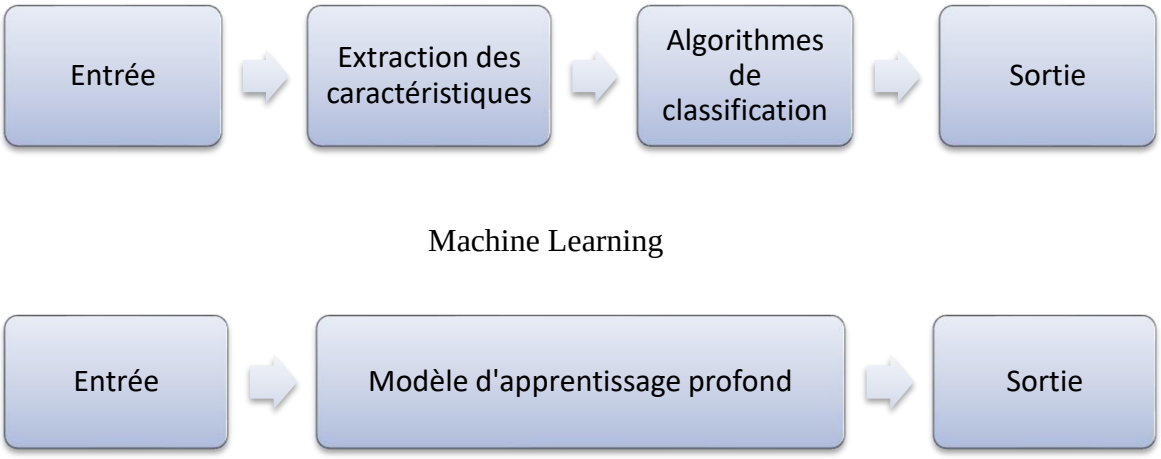


Figure II.6: *Machine learning* Vs *Deep learning*.

Par contre, le *deep learning* est une branche d'apprentissage automatique, il donne la capacité à la machine d'apprendre comment réaliser des tâches de classification directement à partir des données brutes (image, texte ou d'audio).

Le principe du *deep learning* repose sur un apprentissage hiérarchique couche par couche, entre chaque couche interviennent des transformations non-linéaires et chaque couche reçoit en entrée la sortie de la couche précédente.

L'un des atouts du *deep learning* est qu'il remplace la phase d'extraction des caractéristiques qui se fait manuellement dans la technique classique par des algorithmes d'extraction de descripteurs hiérarchiques.

Le *deep learning* a été introduit dans les années 1980, mais ses capacités n'ont été exploitées que récemment à cause de deux points essentiels : premièrement la nécessité de grandes bases de données labellisées pour l'apprentissage et deuxièmement la puissance de calcul considérable nécessaire.

Le terme profond est lié au nombre de couches cachées, contrairement aux réseaux de neurones multicouches qui ne comporte que 2 à 3 couches, les réseaux profonds peuvent avoir jusqu'à une centaine de couches cachées.

Les systèmes de classification à base du *deep learning* ont montré qu'ils peuvent atteindre des taux de reconnaissance exceptionnelle en utilisant de très grandes bases de

données pour l'apprentissage et des modèles de réseaux de neurones artificiels à plusieurs couches cachées.

II.3.1. Domaines d'application du *Deep Learning*

L'importance de l'apprentissage profond découle des résultats très performants obtenus dans divers secteurs, actuellement on trouve que le deep learning a été intégrer dans les nouvelles voitures autonomes pour la détection des panneaux de signalisation, des piétons et des obstacles dangereux sur la route, dans l'aéronautique (drones), dans la téléphonie (capteurs photo intelligents), dans l'identification et la détection des objets à partir des images satellitaires de grandes dimensions ainsi que dans le domaine médical tel que la détection des cellules cancéreuses.

Dans le domaine de la reconnaissance d'activités humaines, le deep learning a permet aux chercheurs de franchir la barrière de la nécessité de conception des descripteurs en rendant cette étape automatisée et aussi d'atteindre des taux de classification inédits.

II.3.2. Les réseaux de neurones à convolution (*Convolutional Neural Networks*)

Les réseaux de neurones artificiels à convolution (CNN ou ConvNet) ou appelé aussi les réseaux de neurones convolutionnels sont le type le plus utilisé dans l'apprentissage profond, à chaque couche du réseau des descripteurs sont extraits , et la complexité des descripteurs augmente en allons dans les couches les plus profondes, dans l'exemple de la figure II.7, la première couche extraite des descripteurs simples tels que la direction et la couleur, la troisième couche qui est la combinaison de la première et la deuxième couche montre des descripteurs plus complexes [60].

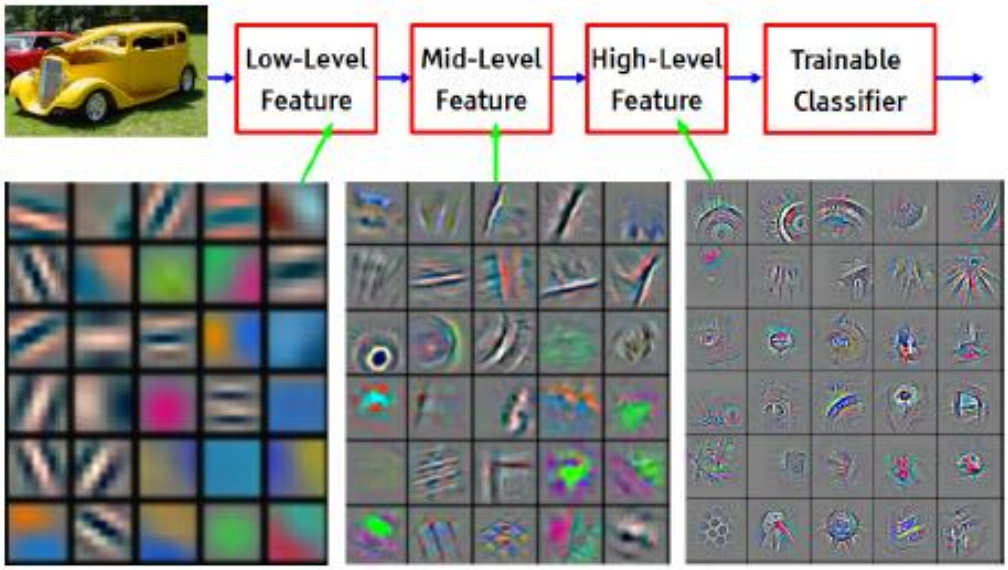


Figure II.7 : Représentations hiérarchiques apprises par un CNN [60].

Dans cette partie, nous allons présenter les différentes couches ainsi que les notions relatives aux réseaux de neurones à convolution.

II.3.2.1. La couche convolutionnelle

Est la couche de base dans un réseau CNN, la sortie de cette couche est la convolution de l'image d'entrée par un filtre de dimension (w*h*d), donc on a besoin de w*h*d neurones avec chacun w*h*d connexions dans un perceptron multicouche. Une seule convolution permet d'obtenir toute une carte d'activation de même taille que l'image d'entrée, l'exemple de la figure II.8 montre l'opération de convolution sur une image de taille 5*5 avec un padding de 1*1 et un filtre de dimension 3*3 [56].

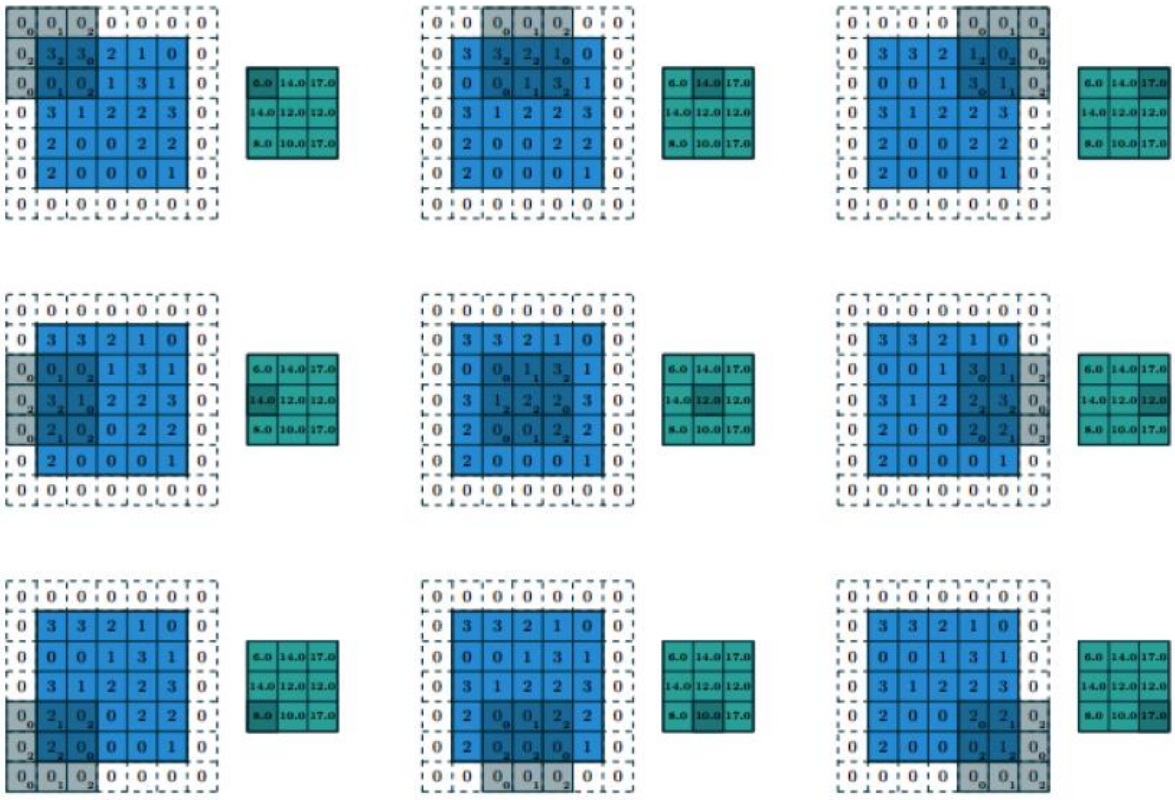


Figure II.8 : Exemple de convolution

La couche de convolution peut être appliquée autant de fois que désirer et obtenir plusieurs cartes d'activation pour une même entrée, c'est le cas du réseau AlexNet [61] qui applique sur l'image d'entrée 96 filtres de convolution de taille (w=11, h=11, d=3) (figure II.9).

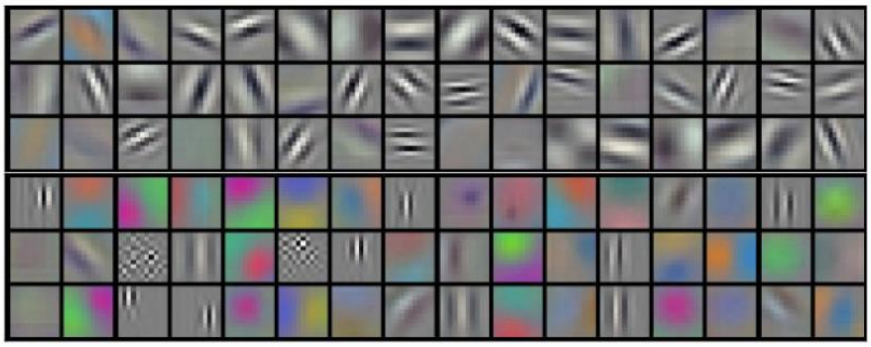


Figure II.9 : Exemples de filtres de convolution : les 96 filtres de la première couche d'AlexNet [61].

Le passage d'une couche de convolution à une autre permet de projeter l'image d'entrée dans un nouvel espace, donc un réseau CNN peut être considéré comme un enchaînement de traitement qui permet de changer la représentation des données [56].

II.3.2.2. Fonction d'activation

Après l'opération de convolution, généralement, on applique une fonction d'activation sur les cartes générées, plusieurs fonctions sont disponibles, mais les plus utilisées dans les réseaux CNN sont le ReLU (*Rectified Linear units*), Leaky ReLU [62].

a- ReLU (Unité de Rectification Linéaire) :

La fonction ReLU est représentée par l'équation suivante [62] :

$$u = \max(0, x) \qquad \text{(II. 3)}$$

C'est la fonction d'activation la plus utilisée dans les réseaux de neurones à convolution, elle remise à zéro toutes les valeurs négatives (figure II.10).

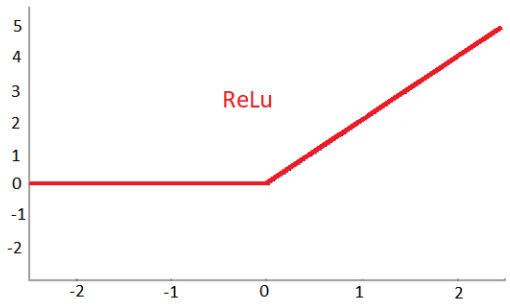


Figure II.10 : La fonction ReLU.

b- Leaky ReLU

Cette fonction est similaire à la fonction ReLU, elle conserve les entrées supérieures à zéro, et multiplie les entrées inferieures à zéro par une constante α (inferieur à 1) (figure II.11). La fonction Leaky ReLU est représentée dans l'équation suivante [62] :

$$u = \max(\alpha x, x) \qquad \text{(II. 4)}$$

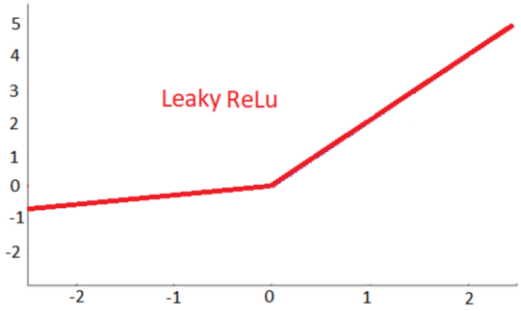


Figure II.11 : La fonction Leaky ReLU (α =0.1)

D'autres fonctions peuvent être aussi utilisées, telles que la fonction sigmoïde et la fonction Sinh.

Dans le domaine du *deep learning*, une couche de convolution est la composition de l'opération de convolution et la fonction d'activation (généralement ReLU) tel que présenté dans la figure II.12.

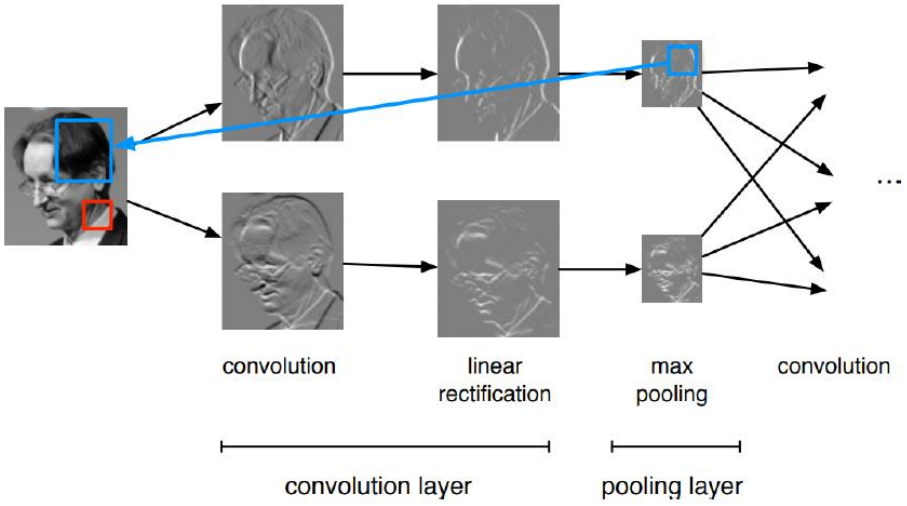


Figure II.12 : Exemple d'une couche de convolution avec deux filtres.

II.3.2.3. Couche de *Pooling*

Dans le but de diminuer la taille des tenseurs des couches intermédiaires, le *pooling* consiste à sélectionner un représentant dans une zone spatiale en fonction d'un critère fixe, le type le plus utilisé est la *Max-pooling* qui prend la valeur maximale dans une zone spatiale de taille 2x2 par exemple (figure II.13), donc on va diviser la taille des tenseurs par 2 ce qui engendre une perte d'information qui reste minime parce que nous gardons le terme le plus représentatif par rapport au filtre précédent, et ainsi d'éviter le sur-apprentissage (*over-fitting*) [56]. D'autres types de *pooling* existe tel que le *average pooling* qui calcule la valeur moyenne dans un filtre de taille MxN.

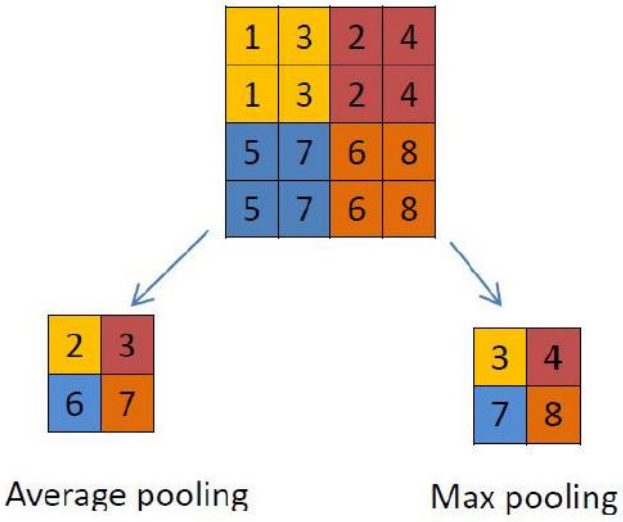


Figure II.13 : Différentes opérations de *pooling* : à gauche, *Average pooling* et à droite le *Max pooling* (un filtre 2x2, stride =2) [63].

II.3.2.4. Couche entièrement connectée (*Fully Connected Layer*)

Appelée aussi couche de produit matriciel, Cette couche effectue un produit matriciel avec son tenseur d'entrée, chaque composante de la sortie est la somme pondérée des paramètres d'une ligne de la matrice de poids avec l'entrée, chaque composante est entièrement connectée au tenseur d'entrée [56]. Cette couche peut être assimilée à une opération de projection, est équivalente à un changement de base entre le tenseur d'entrée et le tenseur de sortie, les deux bases n'ayant pas les mêmes nombres de paramètres.

La couche entièrement connectée projette le tenseur précédent dans un espace de taille égale au nombre de classes [56].

II.3.2.5. La couche (*Softmax*)

Cette couche est précédée toujours par une couche entièrement connectée, son rôle est d'attribuer des probabilités à chaque classe du problème étudié en fonction des données fournis par les couches supérieures du réseau CNN.

II.3.2.6. La couche de classification (*classification layer*)

C'est la dernière couche dans un réseau de neurones à convolution qui traite un problème de classification (dans le cas d'un problème de régression, on trouve une couche de régression),

son rôle est de convertir les probabilités à la sortie de la couche supérieure *Softmax* en classes [56].

Généralement, lorsqu'on parle des réseaux de neurones artificiels, plusieurs acronymes sont utilisés :

Le Batch : est le regroupement de tenseurs 3D (images) dans un nouveau tenseur de dimension 4D ayant pour but d'augmenter la vitesse de traitement.

La normalisation de batch consiste à homogénéiser et normaliser la distribution statistique des données dans un tenseur 4D.

Pour que la normalisation soit efficace, il est nécessaire d'utiliser plusieurs images pour constituer le batch (8, 16, 32, voire 256), mais à la limite de la mémoire disponible. Dans le cas des réseaux très lourds (par rapport aux ressources disponibles), il est nécessaire d'utiliser un batch égal à 1, dans ce cas, il est recommandé de ne pas utiliser la normalisation des batch, parce que cela n'a aucun sens de normaliser une seule image [56].

Le Stride : il spécifie le pas utilisé par l'opérateur de convolution, ce paramètre permet de diminuer la dimension du tenseur de sortie (figure II.14).

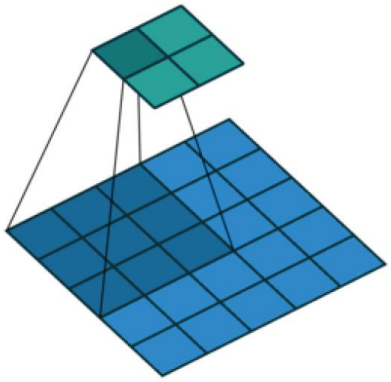


Figure II.14 : Exemple d'un stride=2.

Le Padding : est l'opération d'ajouter des bordures autour de l'image d'entrée dans le but est de garder la taille du tenseur de sortie similaire à celui de l'image originale **(figure II.15)**.

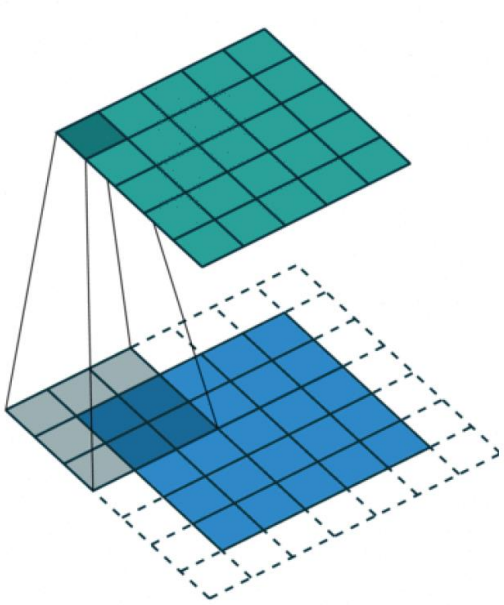


Figure II.15 : Exemple d'un Padding de 1x1.

II.3.3. Conception ! de réseaux de neurones à convolutions

L'apprentissage profond est le fruit de l'évolution des capacités de traitement des machines actuelles (cartes graphiques très performantes et disque de stockage) ainsi que de la disponibilité de bases de données labellisées très larges.

Malgré ça, la conception d'un réseau de neurones à convolution n'est pas une tâche simple, plusieurs contraintes se présentent :

- L'architecture du réseau : aucune règle n'existe pour concevoir un réseau CNN (Nombre de couches, nombre de filtres de convolution, …, etc.).
- Le temps d'apprentissage : qui est très grand dans le cas des réseaux CNN larges, tel que le cas de traitement vidéo (l'apprentissage peut prendre des semaines).
- Un temps de conception très grand à cause du temps d'apprentissage ainsi que des multitudes de simulations à réaliser pour trouver la bonne architecture.
- Le coût : la nécessité de machines de calculs très performantes équipées d'une ou plusieurs cartes graphiques dont le prix peut atteindre quelques milliers de dollars, ainsi que des disques de stockage volumineux pour stocker des bases de données ayant une dimension qui va jusqu'à 10 To.

L'utilisation des réseaux de neurones à convolution dans la résolution d'un problème donné peut se faire de trois manières différentes :

- Conception d'une architecture à partir de zéro
- Utilisation de l'apprentissage par transfert (*transfert learning*)
- Utilisation de l'affinement (*fine-tuning*)

La conception d'un réseau à partir de zéro peut prendre des jours voir des semaines ainsi la nécessité de contourner les contraintes déjà cité précédemment.

La deuxième alternative est l'apprentissage par transfert (*transfert learning*), ici on considère un modèle entrainé sur une base de données pour résoudre un problème donné , on enlève les dernières couches entièrement connectées ainsi que la couche *Softmax*, puis on insère une nouvelle couche entièrement connectée et une nouvelle couche *Softmax*, on gèle les poids des couches supérieures, ensuite on refait l'apprentissage en utilisant la nouvelle base de données pour résoudre le nouveau le problème. On prend par exemple le réseau AlexNet qui a été entrainé sur la base de données d'ImageNet pour la classification des objets, et on va remplacer la dernière couche entièrement connectée et *Softmax*, ensuite on refait l'apprentissage en utilisant la nouvelle base de données qui est constituée des images par exemple biomédicales, image satellite, …etc.

L'avantage de l'apprentissage par transfert est d'utiliser des architectures qui ont prouvé leur efficacité, de garder les cartes de descripteurs complexes entraînées sur des millions d'images et de les adapter à notre problème, ce qui permet de minimiser considérablement le temps d'apprentissage.

Fine-tuning ou l'affinement d'un modèle CNN est la troisième alternative, cette méthode est similaire au transfert learning, mais ici, on peut refaire l'apprentissage de tous ou de quelques couches seulement, donc on peut considérer le *transfert learning* comme un cas particulier du *fine tuning*. Le workflow du *fine tuning* est le suivant (figure II.16) :

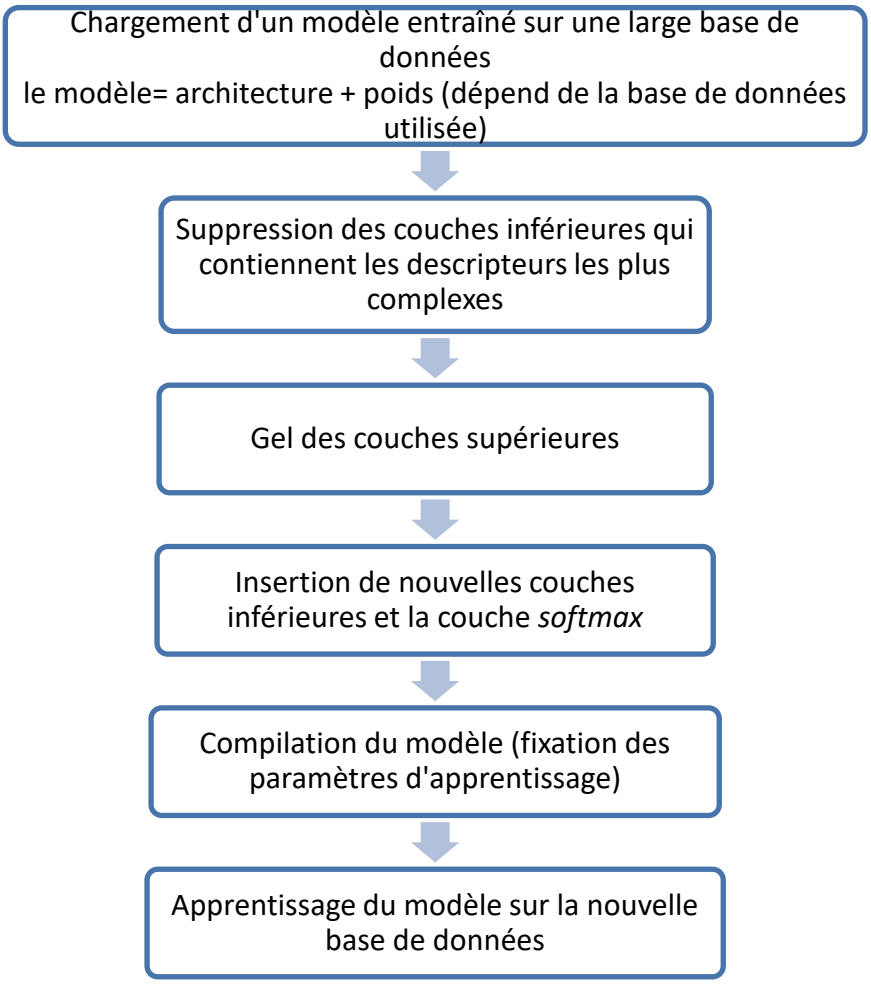


Figure II.16 : Le workflow du fine-tuning.

II.3.4. Bases de données utilisées pour l'apprentissage des réseaux CNN

Comme précisé précédemment dans le workflow du fine-tuning, un modèle est constitué de l'architecture et des poids qui sont les résultats d'apprentissage du réseau sur une base de données, il est donc primordial de bien connaître cette base de données, parce que les performances du modèle dépendent directement de cette dernière, par exemple un modèle qui a été entrainé sur la base de données ImageNet [64] de Google qui contient plus de 15 millions d'images labellisées en 22000 catégories donne de meilleurs résultats que lors de l'utilisation d'autres bases de données moins large tel que COCO [65] qui contient 80000 images pour l'apprentissage et 40000 pour la validation, catégorisées en 80 classes.

D'autres bases de données sont adaptées à des problèmes précis comme la segmentation, tel que la base de données Pascal database [66].

II.3.5. Etat de l'art sur les réseaux de neurones à convolution

Dès le lancement de la compétition *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) de google en 2012, plusieurs réseaux de neurones à convolution ont été proposés dans la littérature dans le but d'obtenir de meilleures performances [67]. La figure II.17 montre les principaux réseaux proposés ainsi que leurs performances, la figure II.18 montre les principaux réseaux en fonction de leurs paramètres.

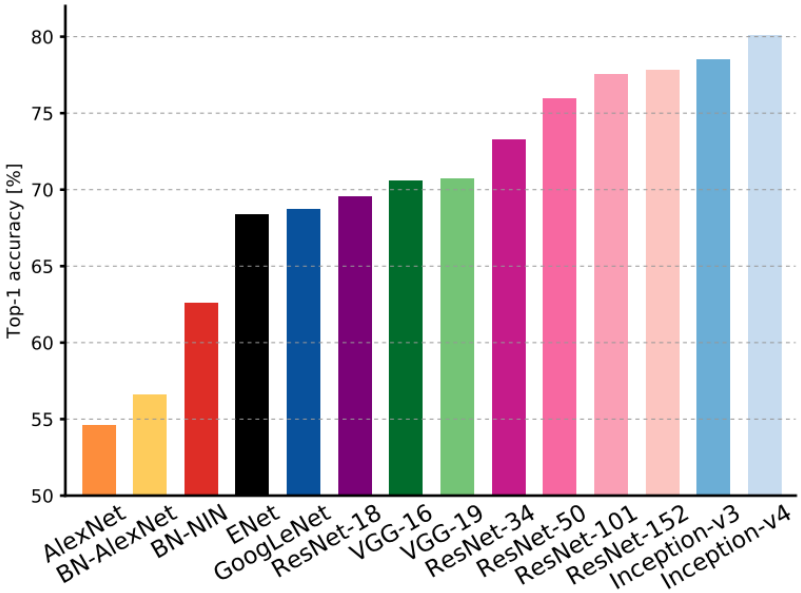


Figure II.17 : Réseaux Vs performances dans l'épreuve Top1 de ImageNet [67].

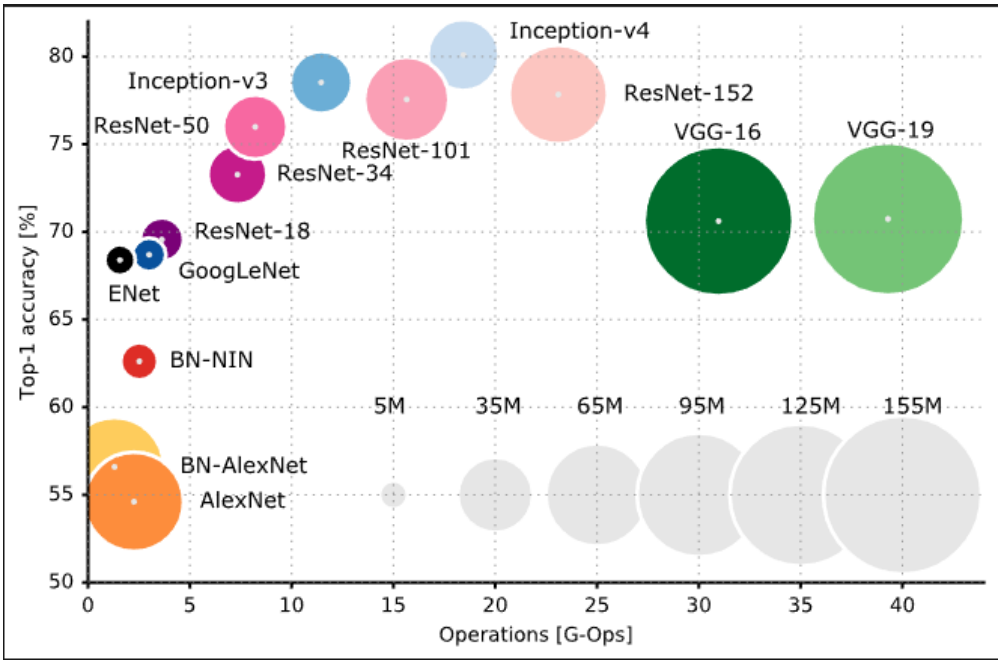


Figure II.18 : Réseaux Vs paramètres dans l'épreuve Top1 de ImageNet [67]

LeNet-5

LeNet-5 est le premier réseau de neurones à convolution qui a été proposé par Lecun Y dans [68] pour la reconnaissance de l'écriture (figure II.19), la première couche du réseau est la couche d'entrée, elle a la même dimension que l'image d'entrée (32x32), elle est obtenue par la convolution de l'image d'entrée par 6 filtres de dimension 5x5 et un stride=1, la deuxième couche est obtenue par l'utilisation d'un *pooling* (*average pooling*) avec un filtre 2x2 qui fait un re-échantillonnage par deux, la troisième couche est une couche de convolution en utilisant 16 filtres de dimension 5x5, la quatrième couche et une couche de *pooling (average pooling)* avec un filtre 2x2, la cinquième couche et une couche entièrement connectée avec 120 descripteurs de taille 1x1, la sixième couche est aussi une couche entièrement connectée avec 84 descripteurs, la dernière couche est une couche de classification *Softmax*.

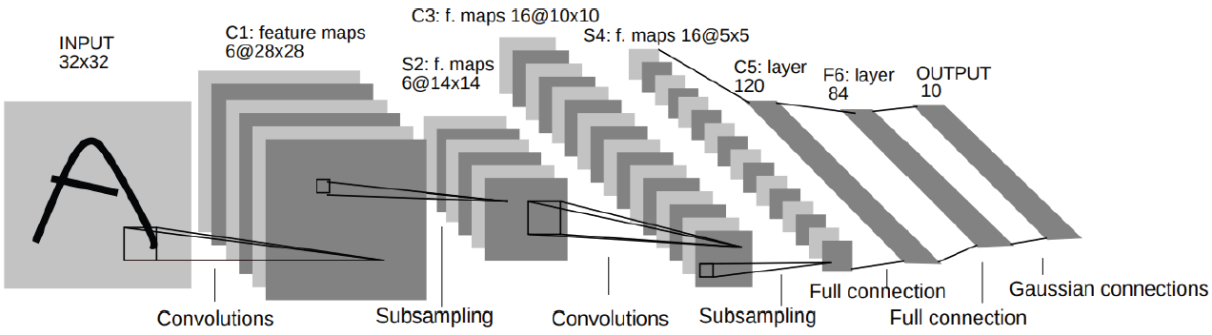


Figure II.19 : Le premier réseau CNN (LeNet-5) [68].

AlexNet

Il est le premier réseau qui a remporté la compétition ImageNet de google en 2012, AlexNet a été proposé dans [61], la première couche dans le réseau est une couche de convolution avec 96 filtres de taille 11x11 et un stride de 4 (figure II.20), ensuite un *Maxpool* est appliqué avec un filtre 3x3 et un stride =2 pour obtenir des cartes de dimension 27x27x96, la deuxième couche est une convolution avec 256 filtres de taille 5x5 et un stride =1 suivi d'un *Maxpool* avec un filtre 3x3 et un stride =2, les couches trois , quatre et Cinq sont des couches de convolution avec 384 filtres pour la troisième et la quatrième et 256 pour la cinquième de taille 3x3, les trois sont suivis d'un *Maxpool* avec un filtre de taille 3x3 et un stride de deux, la sixième couche est une couche entièrement connectée avec 9216 descripteurs, la septième et la huitièmes couches sont aussi des couches entièrement connectées de taille 4096 chacune, la dernière couche est une couche de classification *Softmax* avec 1000 classes possibles. La

fonction d'activation utilisée est un ReLU. Le réseau AlexNet contient 60 millions paramètres et 650000 neurones [61].

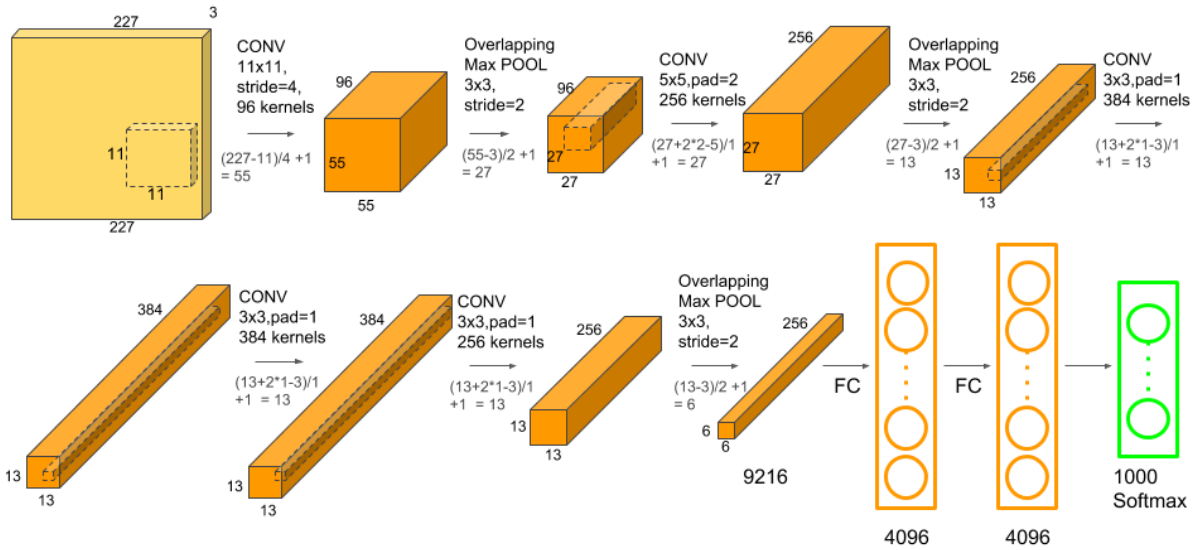


Figure II.20 : Le réseau d'AlexNet.

VGG16 (Visual Geometry Group)

Ce réseau a été proposée par K. Simonyan et A. Zisserman dans leurs papier "*Very Deep Convolutional Networks for Large-Scale Image Recognition*" dans [69], VGG16 a obtenu 92.7% dans l'épreuve de ImageNet, il a été entrainé sur 16 millions d'images pendant des semaines sur une carte graphique NVIDIA Titan Black, le réseau possède 16 couches cachées, et il peut classifier 1000 classes.

La figure II.21 suivante représente l'architecture du réseau VGG16 :

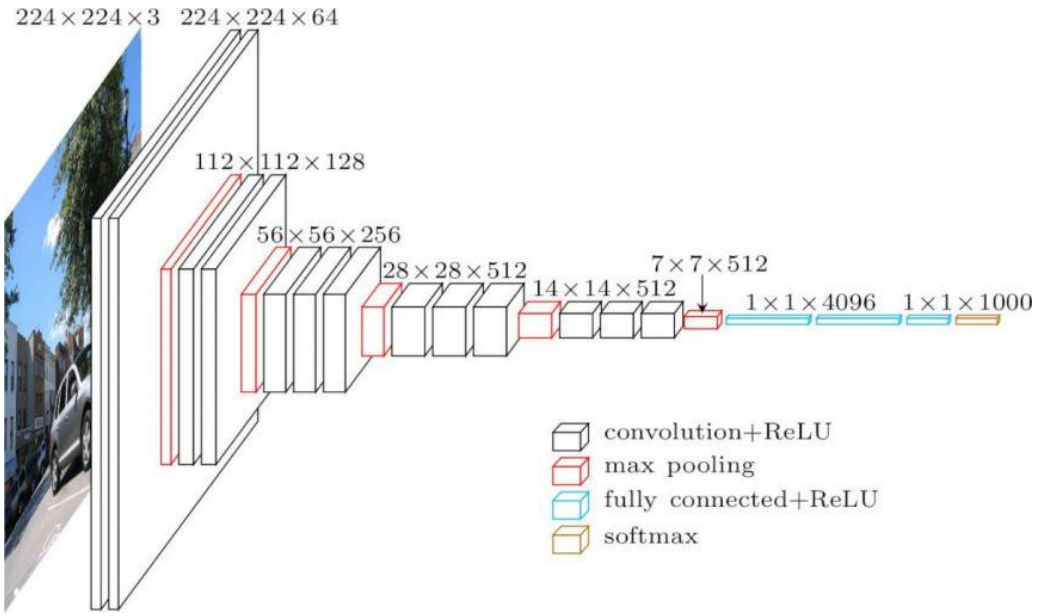


Figure II.21 : VGG16 [69].

VGG19

Proposé dans [69], ce réseau a la même architecture que VGG16, avec trois couches de convolution en plus, donc 16 couches de convolution et 3 couches entièrement connectées, 19 en total (figure II.22).

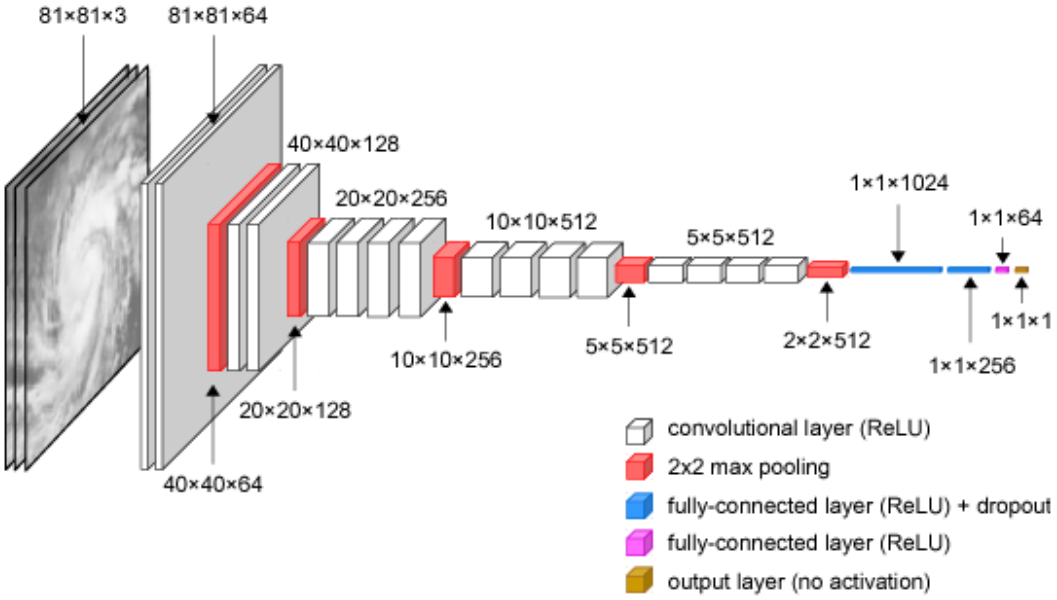


Figure II.22: VGG19 [70].

Cependant, les réseaux VGG soufrent de deux limitations :

- Très lent dans l'apprentissage.
- La taille du réseau est très grande (533 MB pour VGG16 et 574 MB pour VGG19) ce qui rend difficile leur déploiement dans des applications.

D'autres modèles de réseaux CNN ont été proposés dans la littérature, on trouve :

GoogleNet [71] qui a remporté le challenge ILSVRC de 2014, sa contribution était la proposition d'un *module Inception* (figure II.23) basé sur le *global AVG pooling* qui a considérablement réduit le nombre de paramètres (4M contre 60M pour AlexNet et 138M pour VGG), la dernière version de ce réseau est Inception V4[72].

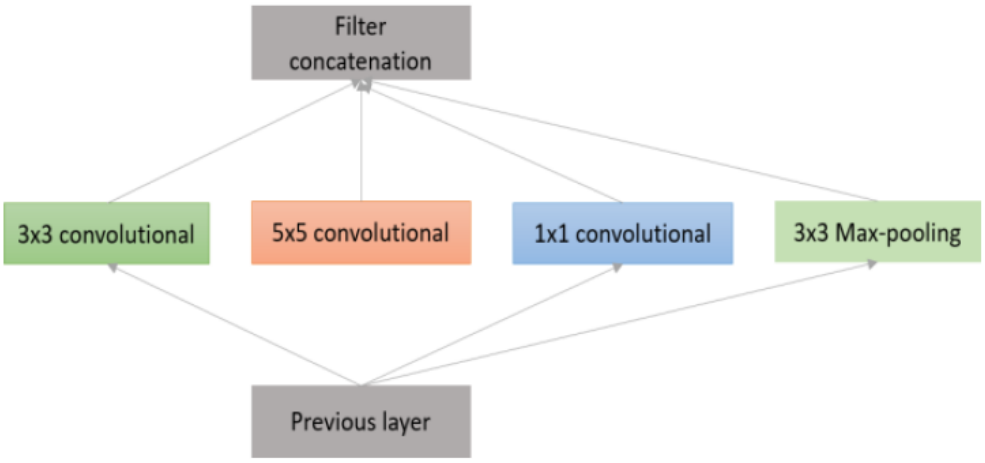


Figure II.23 : La couche Inception proposée par GoogleNet pour réduire le nombre de paramètres [73].

ResNet [74] est le vainqueur du challenge ILSVRC de 2015, la caractéristique de ce réseau est qu'il utilise des sauts de connexions et une forte utilisation du batch de normalisation ainsi que le global *AVG pooling*.

II.4. Les méthodes de reconnaissance d'activités humaines basées sur le Deep Learning

Dans la première partie de ce chapitre, nous avons présenté les notions de base de l'apprentissage profond en commençant par les réseaux de neurones artificiels puis les réseaux de neurones à convolution.

Dans cette deuxième partie, nous allons présenter les méthodes de reconnaissance d'activités humaines qui utilisent l'apprentissage profond.

Plusieurs techniques de reconnaissance d'activités humaines basées sur le deep learning ont été proposées dans la littérature, on peut les classer dans deux catégories selon le type de données utilisées pour l'extraction des descripteurs (figure II.24).

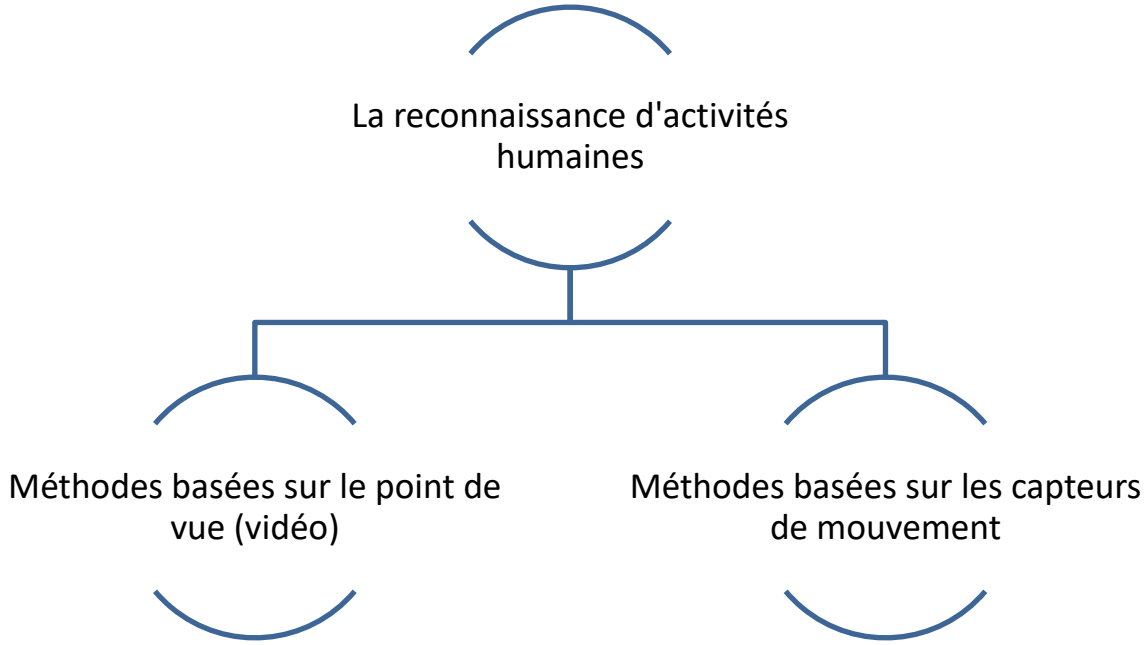


Figure II.24 : Classification des méthodes de reconnaissance d'activités humaines.

Les méthodes basées sur le point de vue sont les méthodes qui utilisent les vidéos comme données d'entrée pour l'extraction des descripteurs, en d'autres termes l'objectif est d'interpréter les activités réalisées par un sujet dans une séquence vidéo.

Par contre, les méthodes qui utilisent les capteurs de mouvement essayent d'interpréter les signaux enregistrés sur un individu par un ou plusieurs capteurs, ces méthodes nécessitent que le sujet porte sur lui des capteurs de mouvement tel qu'un smartphone. Ensuite, utilisent les données enregistrées pour l'extraction des descripteurs.

Dans cette thèse, nous nous intéressons aux méthodes basées sur le point de vue, donc dans ce qui suit nous allons présenter l'état de l'art des techniques de reconnaissance d'activités humaines à partir des scènes vidéo en utilisant l'apprentissage profond.

II.5. Fusion de l'information temps à travers un réseau de neurones profond CNN

Avant de présenter l'état de l'Art des techniques de reconnaissance d'activités humaines, il est important de présenter la notion de la fusion de l'information spatio-temporelle.

L'un des problèmes majeurs dans la reconnaissance d'activités humaines en utilisant les vidéos, est la représentation et l'introduction de l'information temps dans l'étape de l'extraction des vecteurs caractéristiques.

Dans les techniques de reconnaissance d'activités humaines en utilisant les réseaux de neurones à convolution CNN, plusieurs méthodes de fusion pour introduire l'information temporelle ont été proposées [13] (figure II.25) :

L'utilisation de la reconnaissance image par image (*Single frame*) : ici, chaque image est traitée séparément par le réseau CNN, est la décision finale ne concerne que cette image.

La fusion tardive (*Late fusion*) : chaque image de la séquence vidéo (espacée de 15 images) est traitée par le même réseau CNN, est la décision finale qui caractérise l'activité dans la séquence est la combinaison des couches entièrement connectées des images espacées de 15 images.

La fusion précoce (*Early fusion*) : ici un nombre d'images est empilé (combinées dans une seule image) et traité par le réseau CNN, la décision finale est la reconnaissance d'activité englobée dans la séquence vidéo.

La fusion lente (*Slow fusion*) : l'information temps est incorporée dans le résultat final à chaque niveau du réseau de neurones CNN.

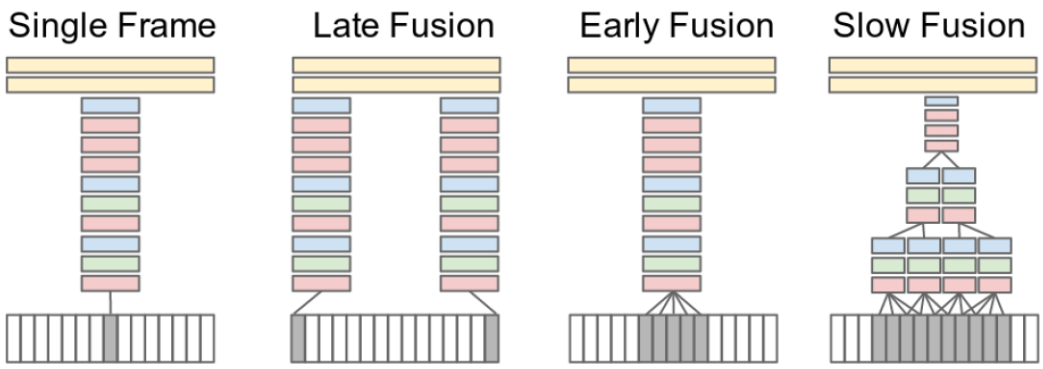


Figure II.25 : Les approches de la fusion pour l'incorporation de l'information temps [13].

II.6. La reconnaissance d'activités humaines en utilisant le descripteur BMI (*Binary Motion Image*)

Cette méthode a été proposée par Tushar D et al. Dans [14], ils proposent de calculer le descripteur BMI (*Binary Motion Image*) à partir des images binaires du sujet après extraction du fond. La figure II.26 montre l'organigramme de la méthode proposée.

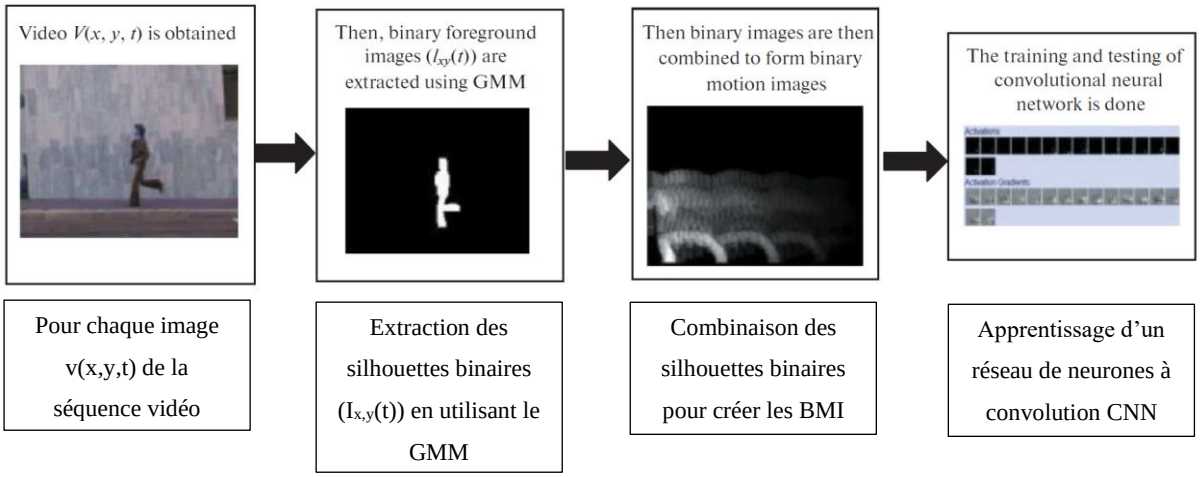


Figure II.26 : Principe de la méthode proposée dans [14].

L'étape de prétraitement qui consiste à l'extraction du fond est réalisée par l'utilisation de la méthode GMM (*Gaussian Mixture Model*) pour la conservation du premier plan, pour l'extraction des descripteurs, les auteurs ont proposé le descripteur BMI (*Binary Motion Image*) qui est la combinaison des images d'avant-plan (*Foreground*) après extraction du fond combiné en une seule image BMI en utilisant l'équation suivante :

$$BMI(x,y) = \sum_{t=1}^{n} f(t) I_{xy}(t) \qquad (II.5)$$

Avec :

$BMI(x,y)$: l'image binaire du mouvement (BMI)

I_{xy} : image d'avant plan.

$f(t)$: est une fonction de pondération qui donne plus de poids aux images récentes.

N : est le nombre total des images dans la séquence.

Les auteurs ont proposé l'utilisation des images en profondeur (*depth images*) par l'utilisation de caméra 3D telle que Microsoft Kinect, ensuite les images de profondeur sont projetées en (x, y, z). A la fin, on obtient 3 descripteurs BMI pour la même séquence vidéo (figure II.27).

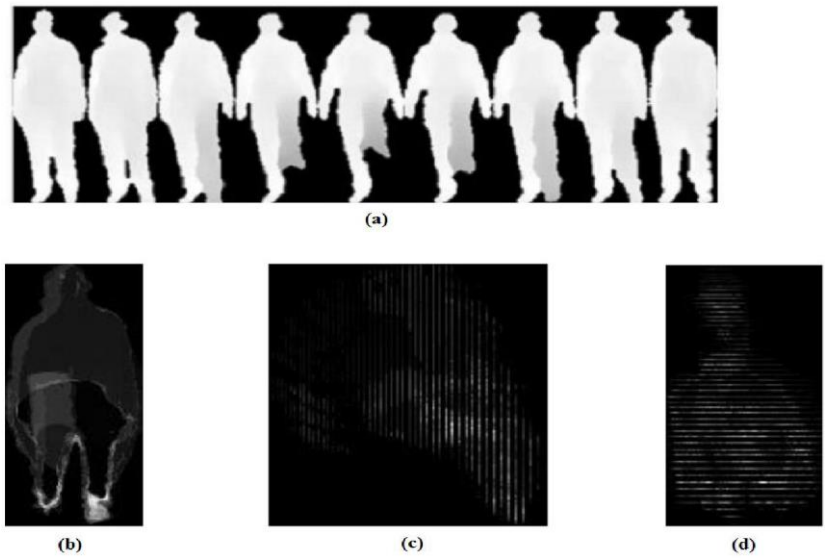


Figure II.27 : a) Image de profondeur, b) Front-View BMI, c) Side-View BMI, d) Top-View BMI [14]

Pour la classification, les auteurs ont proposé l'utilisation des réseaux de neurones à convolution (CNN), il propose d'utiliser l'architecture LeNet-5 de LeCun [68] dans la figure (figure II.19).

Cette méthode a été testée sur deux bases de données, Weizmann et MSR Action3D. Malgré les résultats remarquables, cette approche ne peut pas être utilisée pour la reconnaissance de plusieurs sujets dans la même séquence vidéo, à cause du chevauchement des descripteurs BMI.

II.7. La reconnaissance d'activités humaines en utilisant l'apprentissage profond séquentiel (*Sequential Deep Learning*)

Moez B et al. ont proposé dans [15] une méthode entièrement automatisée, sans aucun prétraitement à priori, composée de deux étapes : la première est l'étape de construction des descripteurs spatio-temporels, la deuxième étape est la classification des activités en utilisant un réseau de neurones récurrents (RNN).

Pour la construction des descripteurs spatio-temporels, les auteurs ont proposé l'architecture de réseaux de neurones à convolution 3D, qu'est une extension directe des réseaux de neurones à convolution 2D conçue pour la prise en charge des vidéos, cette architecture est représentée dans la figure II.28 suivante :

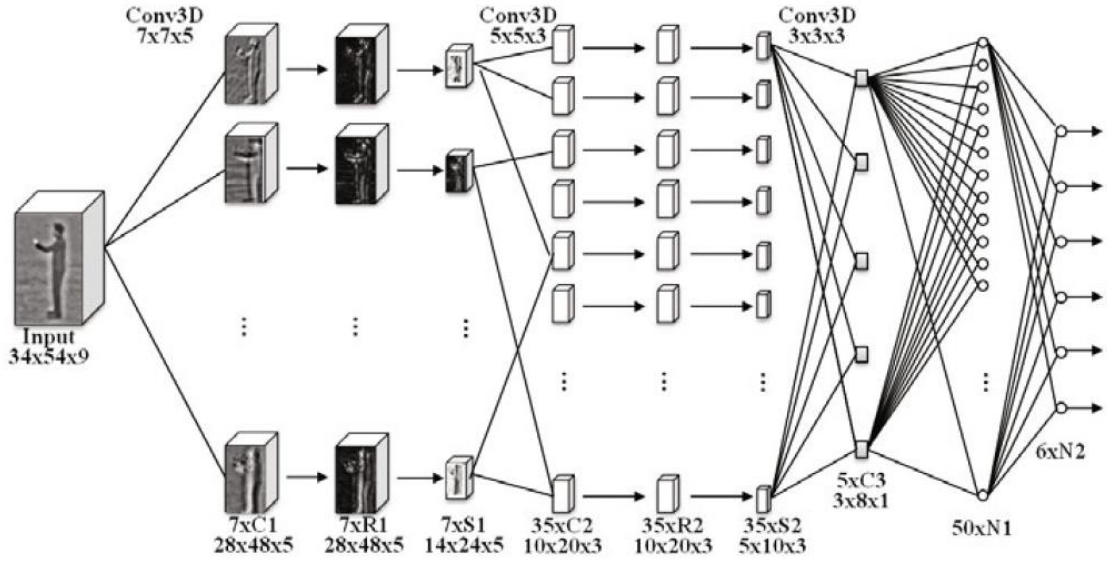


Figure II.28 : Architecture 3D-ConvNet utilisée pour la construction des descripteurs spatio-temporels [15].

Cette architecture est composée de 10 couches, avec deux alternances composées de : une couche de convolution, rectification et sous-échantillonnage (C1, R1, S1) et (C2, R2, S2), suivie d'une troisième couche de convolution et deux couches entièrement connectées N1 et N2 avec en total 17169 paramètres.

Chaque séquence vidéo est décomposée en n sous-séquences, chaque sous-séquence est l'entrée des couches de convolution/rectification/sous-échantillonnage , donc à la sortie on obtient 7 cartes de descripteurs de taille 14x24x5 , ces cartes sont les entrées des deuxièmes couches convolution/rectification/sous-échantillonnage pour obtenir à la sortie 35 cartes de descripteurs de taille 5x10x3, la couche C3 est totalement connectée à la couche précédente S2, donc à la sortie on obtient 5 cartes de descripteurs de taille 3x8x1, donc chaque activité est représentée par un vecteur de descripteurs de taille égale à 120.

Pour la classification des activités, les auteurs ont proposé l'utilisation des réseaux de neurones récurrents RNN du type LSTM (*Long Short-Term Memory*), l'organigramme de la méthode est représenté dans la figure II.29 suivante :

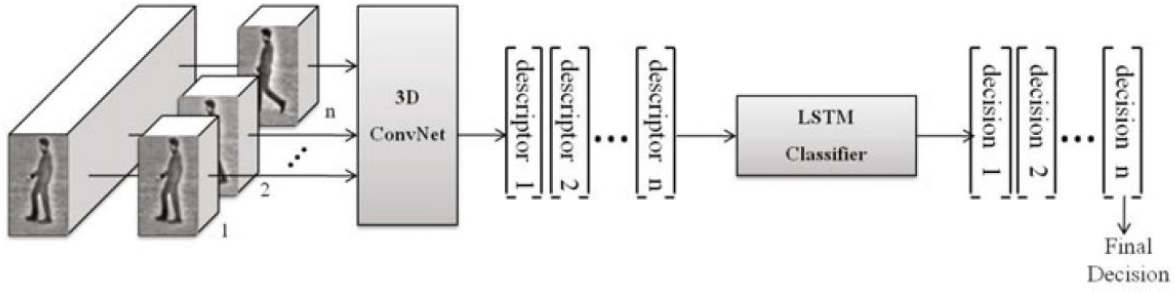


Figure II.29 : Organigramme de la méthode proposée dans [15]

L'entrée du réseau LSTM est les vecteurs caractéristiques de dimension 120 (à la sortie de la couche C3) de chaque sous-séquence. Ce réseau RNN est composé de 50 LSTM caché [15].

II.8. La reconnaissance d'activités humaines en utilisant les cartes de profondeur et les réseaux de neurones à convolution

Pichao W et al. Dans [16] présentent une méthode de reconnaissance d'activités humaines à partir des images de profondeur (*depth images*), ils proposent le WHDMM (*Weighted Hierarchical Depth Motion Maps*) et trois ConvNet fusionnés dans la phase de décision (figure II.30).

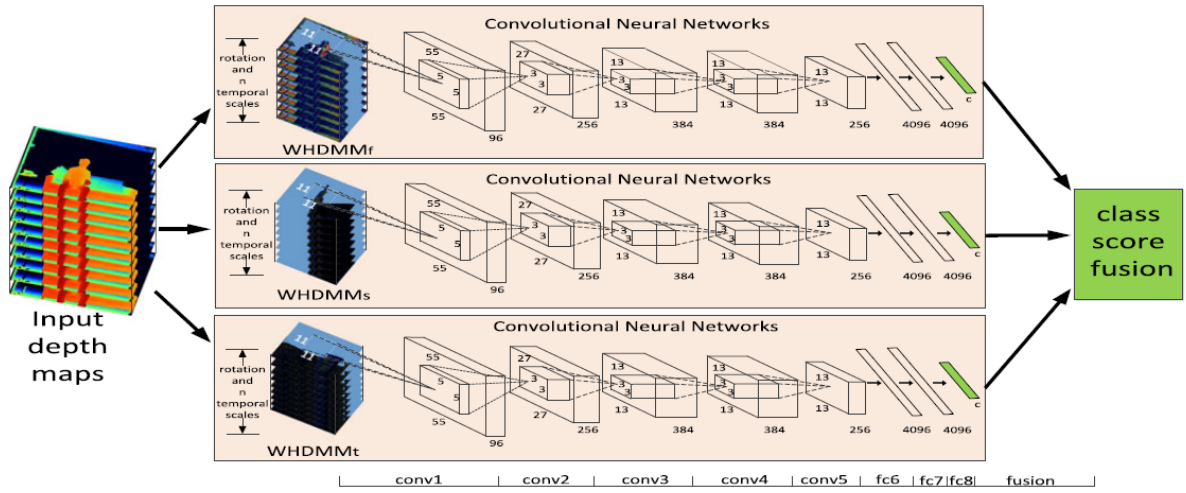


Figure II.30 : Organigramme de la méthode proposée dans [16].

La méthode proposée est composée de trois réseaux ConvNet, la construction des WHDMMs est réalisée à partir des séquences des images de profondeur, chaque réseau

ConvNet reçoit la projection des images de profondeur (WHDMMf, WHDMMs, WHDMMt) sur un plan orthogonal (*front, side et top*).

Pour remédier au problème de la non-disponibilité d'une base de données large (l'apprentissage des réseaux ConvNet nécessite des bases de données très larges), les auteurs proposent deux solutions :

- Générer une base de données plus large à partir des données disponibles par rotation des images de profondeur (figure II.31) pour simuler différentes positions de la caméra (différents points de vue de caméra)
- Utiliser le transfert learning sur les 3 réseaux ConvNet dont l'apprentissage a été réalisé sur la base de données ImageNet de Google.

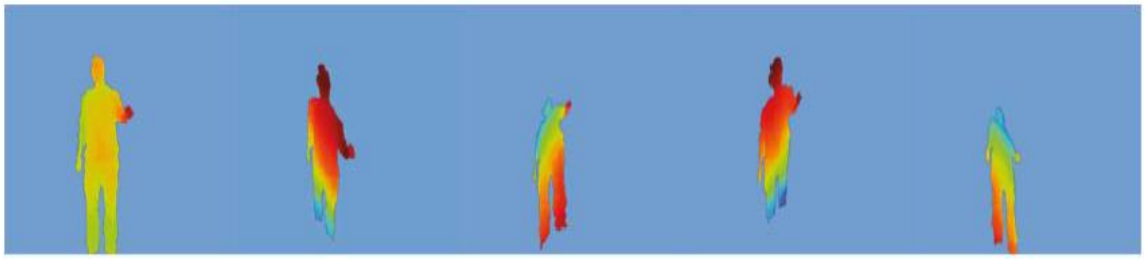


Figure II.31 : Exemples d'image généré par rotation des images en profondeurs [16].

Pour appliquer le transfert learning, les auteurs ont dû transformer le problème de reconnaissance d'activités humaines en un problème de classification des images. Pour ce faire, ils proposent d'utiliser la méthode d'encodage pseudo-couleur (*Pseudocolor Coding*) proposée dans [75] pour la transformation des WHDMM en une seule image (figure II.32).

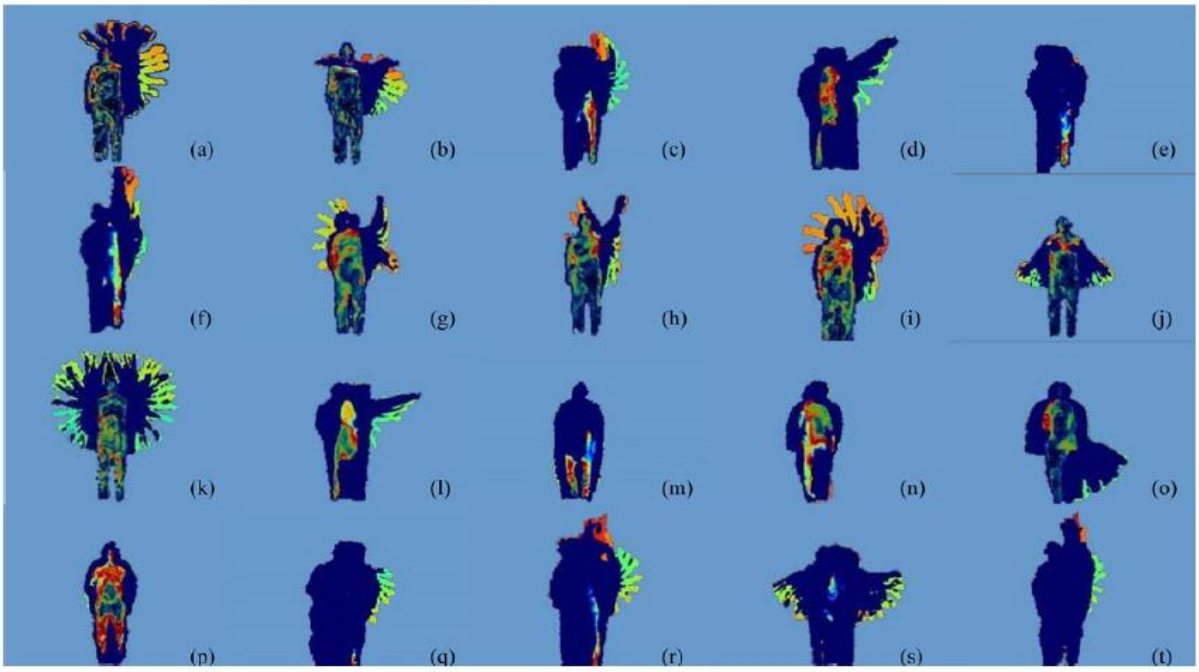


Figure II.32 : Exemple de codage pseudo-couleur des WHDMMs[16].

II.9. La reconnaissance d'activités humaine en exploitant l'apprentissage profond

Dans [17] Pasquale F et al. proposent une méthode de reconnaissance d'activités humaines qui utilise un réseau DBN (*Deep Belief Network*) composé de plusieurs machines de Boltzmann structurées.

Leur méthode est basée sur l'extraction de caractéristiques (ADI (*Average Depth Image*), MHI (*Motion History Image*), et DDI (*Depth Difference Image*) à partir des images en profondeur acquisent par un capteur Kinect, ensuite de les introduire dans un réseau d'apprentissage profond, le principe de la méthode est schématisé sur la figure II.33.

Les auteurs ont proposé d'utiliser une étape d'extraction des caractéristiques plutôt que d'utiliser les données brutes comme entrées du réseau DBN pour augmenter les performances des machines RBM ainsi que pour améliorer le cout de l'apprentissage (temps et machine).

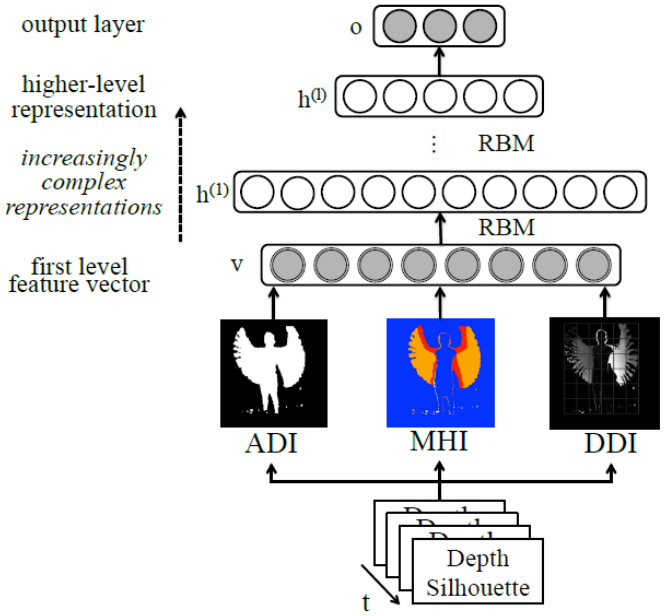


Figure II.33 : Organigramme de la méthode proposée dans [17].

II.10. La classification des vidéos à grande-échelle en utilisant les réseaux de neurones à convolution

Cette méthode a été proposée par Andrej K et al. Dans [13], il s'agit d'une méthode de reconnaissance d'activités humaines multi-résolution par l'utilisation de deux canaux, chaque canal est un réseau de neurones à convolution qui traite les images de même résolution, le principe de la méthode est présenté sur la figure II.34:

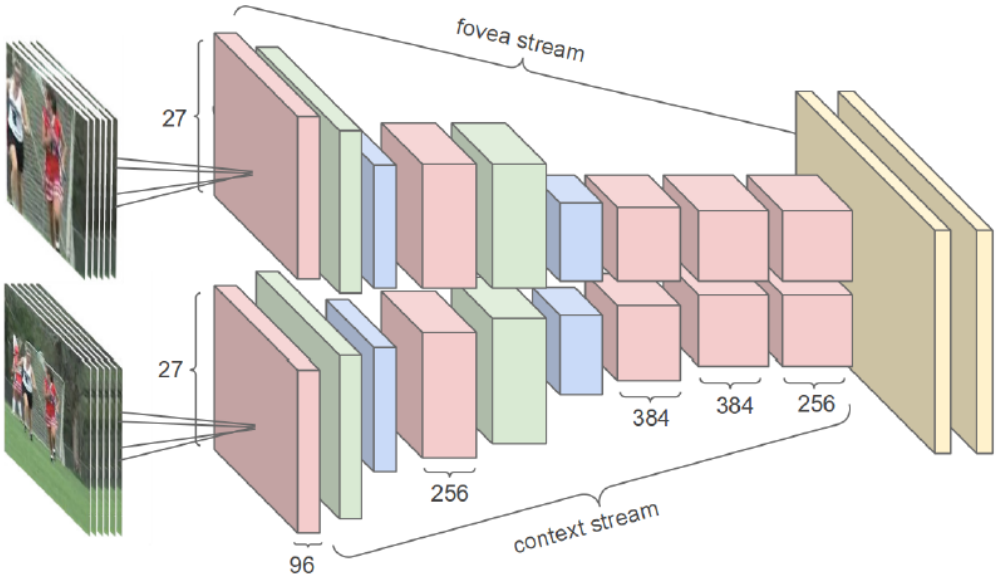


Figure II.34 : Approche proposée dans [13].

Afin de minimiser le temps d'apprentissage et de produire un modèle rapide, les auteurs proposent de travailler sur des images à basse résolution.

Prenant l'exemple des images de résolution 178x178, ils proposent le canal *context stream* qui reçoit des images rééchantillonnées par deux (2) pour générer des images de résolution 89x89, le deuxième canal appelé *fovea stream* reçoit des images centrées sur les sujets de taille 89x89 avec la résolution d'origine, pour obtenir à la fin une dimension divisée par la deux par rapport aux données d'origine de dimension 178x178.

La figure II.35 représente les descripteurs de chaque canal de la méthode proposée :

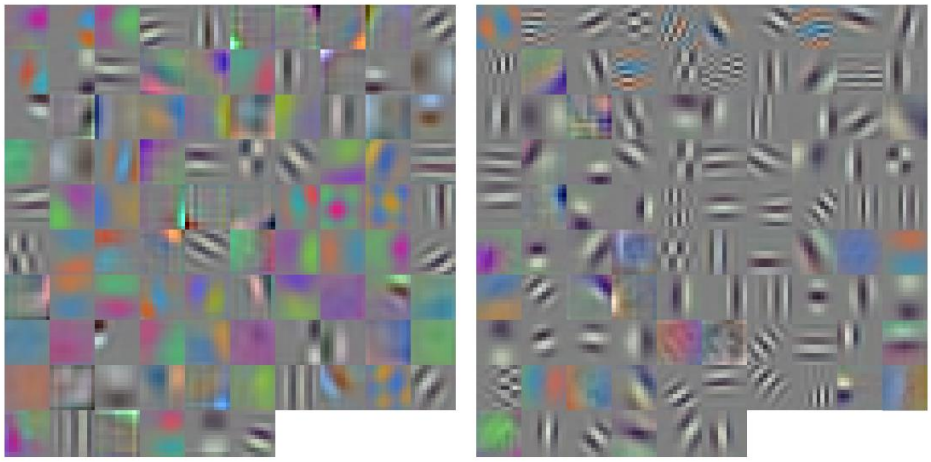


Figure II.35 : Descripteurs de chaque canal de la méthode proposée dans [13]

La figure II.35 montre que la canal *Fovea stream* a permis l'extraction des hautes fréquences ainsi que des niveaux de gris, par contre le canal *Context tream* a permis l'extraction des couleurs et des basses fréquences

II.11. Réseau de neurones à convolution à deux canaux pour la reconnaissance d'activités humaines (*Two-Stream*)

Cette méthode a été proposée par Karen Simonyan et Andrew Zisserman dans [18], les auteurs ont proposé une architecture basée sur deux canaux, un canal spatial et un canal temporel.

L'information spatiale est extraite à partir des images directes de la séquence vidéo, elle contient des informations sur la scène et les objets dans la vidéo, alors que l'information temporelle est extraite à partir d'un ensemble d'images de la séquence vidéo, elle contient des informations sur le mouvement du sujet ainsi que de la caméra.

La figure II.36 donne l'organigramme de la méthode proposée dans [18] :

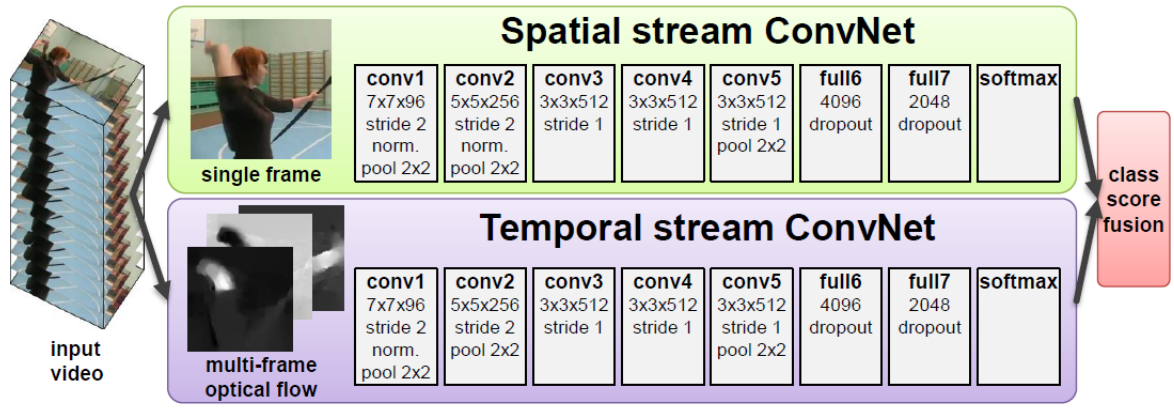


Figure II.36 : Organigramme de la méthode *two stream* [18]

Le premier canal appelé *spacial stream ConvNet* est un réseau de neurones à convolution dont l'entrée est les images individuelles de la séquence vidéo.

Le deuxième canal appelé *Temporal stream ConvNet* lui aussi est un réseau de neurones à convolution dont l'entrée est l'empilement du flux optique calculé entre plusieurs images consécutives de la séquence vidéo, la figure II.37 montre un exemple de calcul du flux optique.

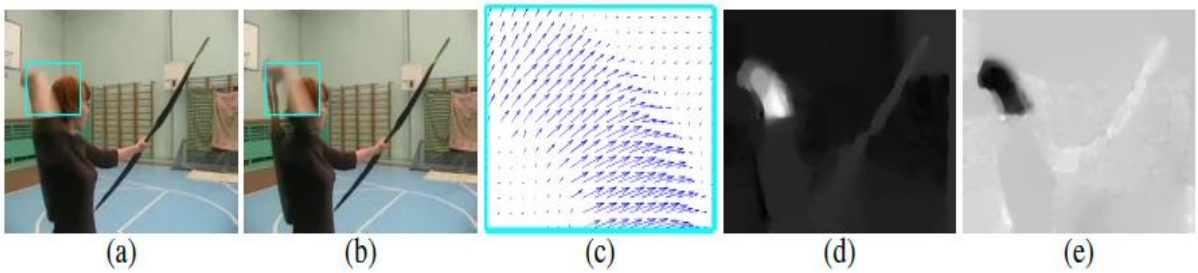


Figure II.37 : Exemple du flux optique : a et b) Deux images successives, c) Le flux optique dans la zone en bleu, d) La composante horizontale du flux optique. E) La composante verticale du flux optique [18].

II.12. Synthèse

L'apprentissage profond ne date pas d'aujourd'hui, mais l'exploitation de sa capacité n'a été possible que récemment grâce à deux points essentiels, l'évolution des unités de calcul (cartes graphiques et disques de stockage) et la disponibilité de larges bases de données labellisées.

L'avantage du *deep learning* par rapport aux autres méthodes du *machine learning* est sa capacité d'auto-extraction et classification à partir des données brutes. Le premier domaine qui a bénéficié de l'apprentissage profond est le domaine de reconnaissance d'objets, dans le cadre du concours *ILSVRC* de Google, plusieurs architectures de réseaux ont été présentées tels que AlexNet, VGG16, VGG19 et GoogleNet. Par la suite, l'apprentissage profond a été introduit dans de nombreux domaines liés au traitement de données en général et à la vision par ordinateur en particulier.

Récemment, plusieurs techniques de reconnaissances d'activités humaines basées sur l'apprentissage profond ont été proposées dans la littérature. Le challenge dans ce domaine réside dans l'extraction de l'information temporelle et sa fusion avec l'information spatiale. Plusieurs protocoles de fusion ont été introduits, le protocole *Single frame* est adapté pour la reconnaissance image par image sans aucune introduction de l'information temporelle, par contre le protocole *Late fusion* permet une fusion au niveau de l'étape de décision des résultats de reconnaissance des images séparées par un intervalle de temps de (15 mages). Le protocole *Early fusion* est basé sur la combinaison de plusieurs image consécutives pour la création d'une image empilée qui contient l'information spatio-temporelle, par cotre le protocole *Slow fusion* permet la fusion de l'information spatio-temporelle à chaque étape du réseau, les trois derniers protocoles sont basés sur la fusion de plusieurs images consécutives donc leur utilisation pour la reconnaissance d'activités image par image en temps réel est impossible, d'autre part, le protocole *Late fusion* est incapable de reconnaitre les activités rapides dont la réalisation peut se faire dans un délai inférieur à 15 images ou dont la variation temporelle est rapide. En plus, le protocole *Slow fusion* nécessite un réseau CNN complexe pour la fusion de l'information temporelle à chaque niveau du réseau.

Dans [14], Tushar et al. Ont proposé une méthode de reconnaissance d'activités humaines basée sur le protocole *Early fusion* en utilisant le descripteur BMI (*Binary Motion Image*) comme entrée au réseau de neurones à convolution LeNet-5. L'inconvénient de cette méthode réside dans le descripteur BMI utilisé, qui est calculé à partir des silhouettes combinées

sur toute l'image. Cela rend l'utilisation de cette méthode pour des applications en temps réel impossible, ainsi que pour la reconnaissance de plusieurs personnes (plusieurs activités) en même temps.

Moez et al. Dans [15], ont proposé une approche basée sur deux types de réseaux, ils proposent d'utiliser les réseaux de neurones à convolution en 3D (3D-CNN) pour l'extraction des caractéristiques et les réseaux de neurones récurrents (RNN) pour la classification des activités. Le fait d'utiliser deux types de réseaux de neurones rend la méthode coûteuse en termes de temps et ressources.

D'autres approches basées sur les images de profondeur ont été proposées aussi, le plus intéressant est le travail réalisé par Pichao et al. Dans [16], ils proposent un système basé sur trois canaux, chaque canal est un réseau de neurones à convolution qui traite les images de profondeur sur un plan orthogonal.

Dans [75], Pascale F et al. Proposent une méthode de reconnaissance d'activités humaines en utilisant les réseaux DBN (*Deep belief Network*) composés de plusieurs machines de Boltzmann, l'entrée de ce réseau est les descripteurs ADI (*Average Depth Image*), MHI (*Motion History Image*), et DDI (*Depth Difference Image*)), comme pour les techniques mentionnées précédemment, cette technique est basée sur l'extraction de descripteurs calculer sur toute l'image , ce qui rend impossible l'utilisation de cette technique pour la reconnaissance d'activités en temps réel.

La technique Two Stream est une des techniques les plus connues dans la littérature, présentées dans [18], l'auteur propose une méthode basée sur deux canaux, dans le premier canal, il utilise un réseau CNN dont l'entrée correspond à des images de la séquence vidéo, le rôle de ce canal est l'extraction de l'information spatiale. Le deuxième canal utilise le même réseau CNN, mais l'entrée dans ce cas correspond à des images du flux optique. Ce canal fait l'extraction de l'information temporelle, à la fin un protocole de fusion au niveau de la phase de classification est appliqué pour la reconnaissance d'activités.

De la même façon, dans [13], Andrej et al. Proposent une méthode basée sur deux canaux, le premier canal utilise un réseau de neurones à convolution dont les 'entrées correspondent aux images de la séquence vidéo divisées en 2, alors que le deuxième canal utilise un autre réseau CNN dont les entrés sont les images centrées sur le sujet.

Les deux dernières techniques sont coûteuses en termes de ressources et de temps, leur utilisation dans la reconnaissance en temps réel est impossible, parce qu'elles nécessitent le calcul des deux canaux séparément avant de donner la décision finale après la fusion.

Dans cette thèse, nous proposons une nouvelle technique de reconnaissance d'activités humaines utilisant l'apprentissage profond, notre méthode est simple par ce qu'elle est basée sur le transfert d'apprentissage du réseau YOLO et sur l'utilisation d'un seul canal. Contrairement aux techniques de l'état de l'art présentées précédemment, notre technique peut être utilisée pour la reconnaissance des activités en temps réel. Elle est utilisable pour la reconnaissance des activités dans des séquences vidéo via l'utilisation d'un protocole de fusion proposé au niveau de la décision. Notre protocole utilise la fusion *Late fusion*, cependant, pour remédier au problème de ce dernier nous proposons la fusion de toutes les images de la séquence vidéo. En profitant des avantages du réseau YOLO, notre méthode est efficace, rapide et surpasse les techniques présentées dans la littérature, comme nous les verrons dans les chapitres suivants.

II.13. Conclusion

Dans ce chapitre, nous avons présenté l'apprentissage profond ainsi que les différentes techniques de référence présentées dans la littérature dans le domaine de la reconnaissance d'activités humaines en utilisant ce dernier.

Dans la partie suivante, nous allons proposer trois méthodes de reconnaissance d'activités humaines et nous allons réaliser plusieurs études comparatives avec les techniques de l'état de l'art.

Partie 2
« Contributions »

Chapitre III

Reconnaissance d'Activités Humaines en utilisant la DCT

« Innover, c'est savoir abandonner des milliers de bonnes idées. »

Steve Jobs
Homme d'affaire, Informaticien, Inventeur (1955 - 2011)

III.1. Introduction

La reconnaissance d'activités humaines en temps réel est un challenge dans le domaine de la vision par ordinateur, dont le but est d'identifier les activités réalisées par des sujets à partir des flux vidéo reçus des caméras. Cela est très important vue la variété des champs d'applications d'un tel système.

Plusieurs méthodes de reconnaissance d'activités humaines basées sur les transformées ont été proposées dans la littérature, on peut citer les travaux de Kamuri et al. [76] qui ont proposé d'utiliser la transformée de Fourier en blocs comme descripteurs. De même, dans [77] Tasweer et al. ont proposé d'utiliser la transformée cosinus discrète à fenêtres appliquée sur les images MHI (*Motion history images*), comme caractéristiques. On peut aussi citer les travaux réalisés par Hafiz et al. [78] qui ont proposé l'utilisation de la transformée de Fourier discrète sur une séquence vidéo, puis une analyse des composantes principales PCA (*Principal Componenet analysis*) pour réduire la dimension et optimiser les descripteurs.

Dans ce chapitre, nous proposons une nouvelle méthode de reconnaissance d'activités humaines en utilisant la transformée en cosinus discrète DCT (*Discret Cosinus Transform*), nous présenterons deux variantes de cette technique, la première a pour but la reconnaissance des activités dans les séquences vidéo. Dans la deuxième variante, nous proposons une nouvelle technique pour la reconnaissance des activités image par image et en temps réel.

Avant d'entamer la description de la méthode proposée et la discussion des résultats expérimentaux, il est important de présenter les bases de données ainsi que les critères d'évaluation utilisés dans le domaine de reconnaissance d'activités humaines.

III.2. Description des bases de données utilisées dans le domaine de la reconnaissance d'activités humaines

III.2.1. La base de données de Weizmann

La base de données de Weizmann a été introduite pour la première fois par Blank et al. [5], elle est composée de 10 activités : Run, Walk, Skip, Jump, Pjump, Side, « Wave using one hand », « Wave using two hands et Bend », chaque activité est réalisée par 9 personnes donc 90 vidéos au total. Les vidéos ont une dimension de 180x144.

Les défis dans la base de données de Weizmann sont : les vidéos sont de résolution médiocre, variation des personnes et vêtement.

Des échantillons de la base de données de Weizmann sont représentés dans la figure III.1 suivante :

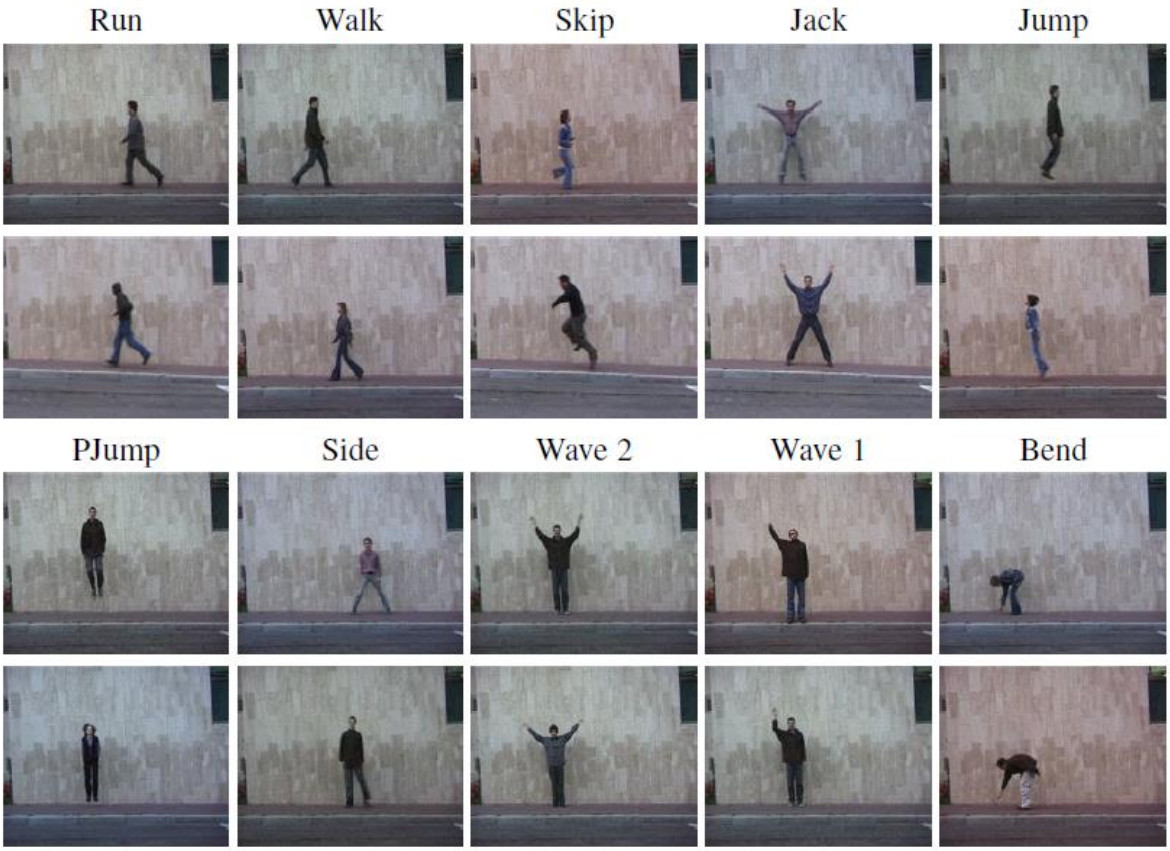


Figure III.1 : Echantillons de la base de données de Weizmann [5].

III.2.2. La base de données Keck Gesture Dataset

Cette base de données a été présentée dans [79], elle contient 14 différentes actions (Turn left, turn right, Att-left, Att-right, Att-both, Stop left, Stop right, Stop both, Flap, Start, come near, Close Dis, speed up, go back) réalisées par 3 personnes, avec une résolution de 640x480. Le challenge avec cette base de données est sa dimension, elle est constituée de seulement 42 activités.

La figure III.2 montre des échantillons de la base de données Keck Gesture Dataset.

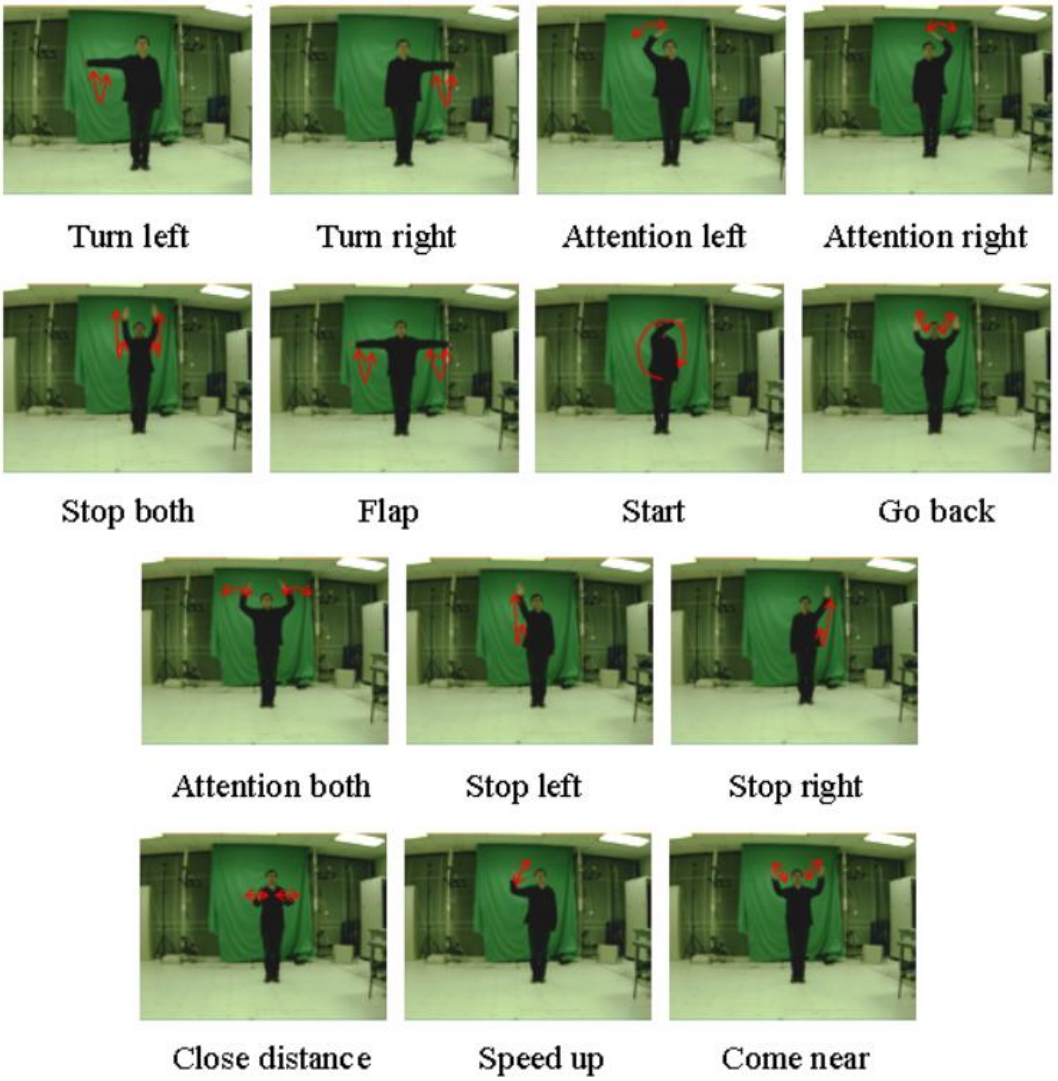


Figure III.2 : Echantillons de la base de données Keck Gesture dataset [79].

III.2.3. La base de données KTH

La base de données KTH est la plus utilisée dans le domaine de la reconnaissance d'activités humaines, elle a été introduite par Schuldt et al. [45], elle est constituée de 6 activités (walking, running, jogging, boxing, hand waving et clapping), chaque action est réalisée par 25 personnes, chaque personne réalise l'action selon 4 scénarios : outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), and indoors (s4).

Les séquences vidéo ont une résolution de 160x120 avec 25 images par seconde, au total la base de données contient 600 vidéos.

Les défis de la base de données KTH sont : vidéos de résolution faible, variations d'échelle, variation des vêtements, changement de luminosité, présence d'ombres, variation de personnes, différents scénarios, variation de la vitesse des actions.

La figure III.3 montre des échantillons de la base de données KTH.

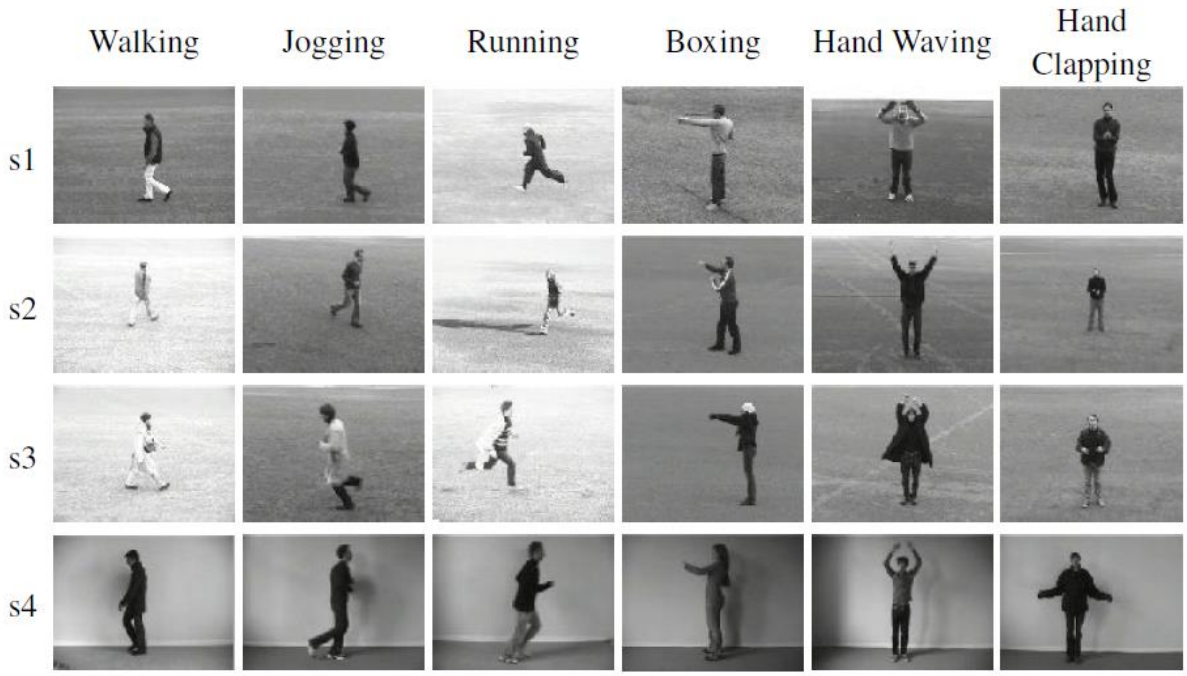


Figure III.3 : Echantillons de la base de données KTH [45].

III.3. Critères d'évaluation

Pour l'évaluation des performances des différentes méthodes de reconnaissance d'activités humaines, et pour permettre de réaliser des études comparatives entre les méthodes proposées et les méthodes de l'état de l'art, nous allons utiliser trois critères d'évaluation : le taux de reconnaissance, la courbe ROC (ROC Curve) et la matrice de confusion.

III.3.1. Taux de reconnaissance

Le taux de reconnaissance ou le taux de classification est le critère de base dans le domaine de la classification en général et particulièrement dans le domaine de la reconnaissance d'activités humaines, c'est le rapport du nombre d'activités dont la classification est correcte sur le nombre total des activités dans l'ensemble de test :

$$Taux\ de\ reconnaissance = \frac{Nombre\ d'activités\ bien\ classées}{Nombre\ total\ des\ activités\ dans\ l'ensemble\ de\ test}$$

III.3.2. Courbe ROC (Roc Curve)

La fonction d'efficacité de récepteur en anglais (*Receiver Operating Characteristic*) ou appelé aussi la courbe sensibilité/spécificité est une mesure de performances d'un modèle de

classification binaire. La courbe ROC permet de mesurer à quel point le modèle est capable de faire la différence entre classes. La figure III.4 montre un exemple de la courbe ROC.

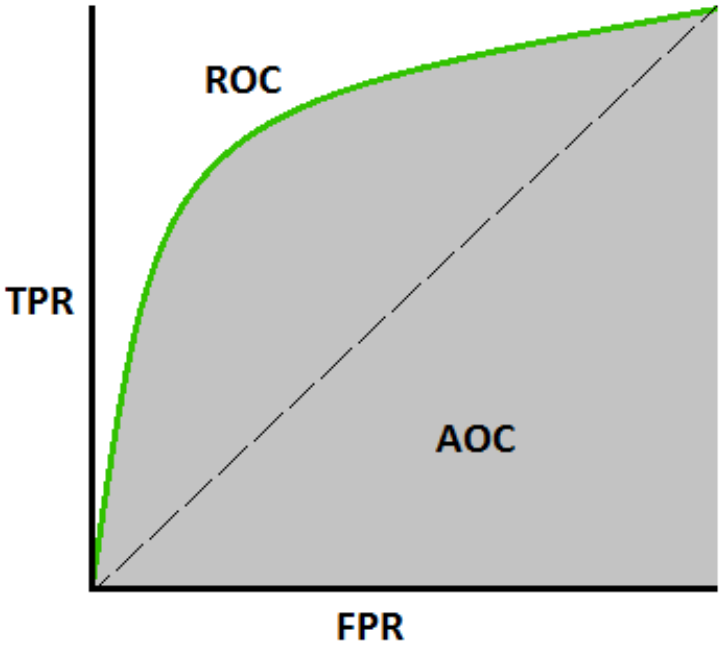


Figure III.4 : Courbe ROC.

L'axe y correspond à la sensibilité (TPR) elle est calculée par l'équation suivante :

$$TPR\ (True\ positive\ rate) = \frac{TP}{TP + FN} \qquad \text{(III. 1)}$$

L'axe x est le FPR $(false\ positive\ rate)$, il est calculé selon l'équation suivante :

$$(Spécificité) = \frac{TN}{TN + FP} \qquad \text{(III. 2)}$$

$$FPR\ = 1 - spécificité = \frac{FP}{TN + FP} \qquad \text{(III. 3)}$$

Avec :

TP : taux des vrais positifs (les cas où la prédiction est positive, et où la valeur réelle est effectivement positive)

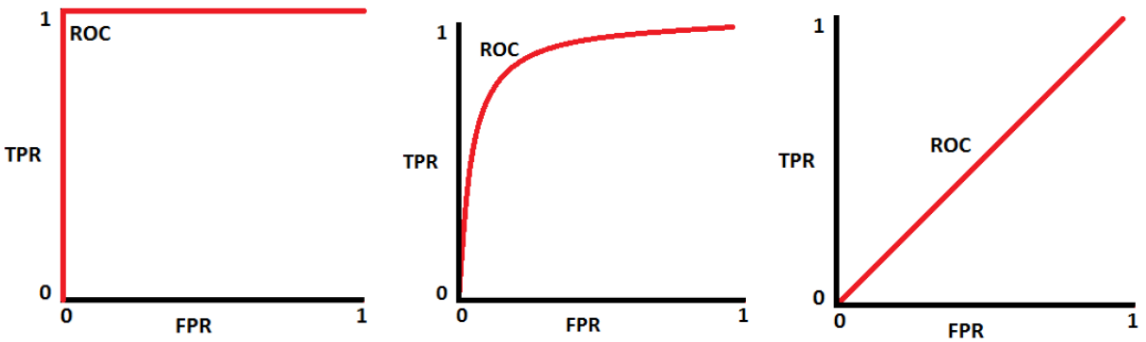
FN : taux des faux négatives (les cas où la prédiction est négative, mais où la valeur réelle est positive)

TN : taux des vraies négatives (les cas où la prédiction est négative, et où la valeur réelle est effectivement négative)

FP : taux des faux positive (les cas où la prédiction est positive, mais où la valeur réelle est négative)

La courbe ROC est interprétée comme suit (figure III.5) :

- La figure III.5a montre le cas d'un modèle de classification idéale où le modèle délivre un taux de reconnaissance de 100%
- La figure III.5b montre le cas d'un classificateur qui peut donner des bons taux de classification.
- La figure III.5c montre le cas le plus défavorable d'un modèle qui donne des résultats aléatoires.



a b c

Figure III.5 : Interprétation du ROC Curve.

III.3.3. Matrice de confusion

La matrice de confusion est un outil d'évaluation des performances de classification d'un modèle. Elle donne un résumé des résultats de classification, les prédictions correctes et incorrectes sont réparties par classes, ce qui permet de comparer les résultats entre classes.

La matrice de confusion permet de visualiser le comportement du modèle de classification, erreurs commises et type d'erreurs. Le principe de cette matrice est représenté sur la figure III.6.

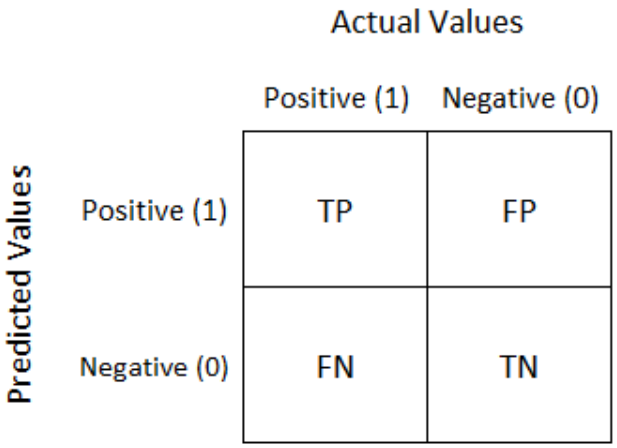


Figure III.6 : Matrice de confusion

III.4. Rappel sur la transformée en cosinus discrète (DCT)

La transformée en cosinus discrète est un outil très populaire dans le domaine du traitement du signal, elle permet la concentration de la majorité de l'énergie d'un signal (image) dans peu de coefficients. Dans notre approche, nous avons essayé de tirer profil de cette propriété afin d'extraire des descripteurs des images issues d'un flux vidéo.

La transformée 1-D DCT de longueur N est donnée par l'équation suivante [80] :

$$\boldsymbol{C}(u) = \alpha(u) \sum_{x=0}^{N-1} f(x) cos\left[\frac{\pi(2x+1)u}{2N}\right], \qquad \text{(III. 4)}$$

avec u $=$ 0,1,2, … , N $-$ 1. La transformée inverse est donnée par :

$$f(x) = \sum_{u=0}^{N-1} \alpha(u) C(u) cos\left[\frac{\pi(2x+1)u}{2N}\right], \qquad \text{(III. 5)}$$

Avec x $=$ 0,1,2, … , N $-$ 1. Dans les équations (1) et (2), α(u) est défini par :

$$\alpha(u) = \begin{cases} \frac{1}{\sqrt{N}}, & u = 0 \\ \sqrt{\frac{2}{N}}, & u \neq 0. \end{cases} \qquad \text{(III. 6)}$$

La transformée en cosinus discrète est définie par les coefficients des basses fréquences (*Details Coefficient*) qui concentrent une grande partie de l'énergie du signal, et les coefficients hautes fréquence (*Approximation Coefficient*) qui informent sur les détails du signal.

La transformée en cosinus discrète 2D est l'extension de la version 1-D DCT, elle est représentée par l'équation suivante [80] :

$$\boldsymbol{C(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) cos\left[\frac{\pi(2x+1)u}{2N}\right] cos\left[\frac{\pi(2y+1)v}{2N}\right]}, \text{(III. 7)}$$

Avec $u, v = 0,1,2, \dots, N-1$ et α(u) and α(v) sont définis dans (3), la transformée inverse est donnée par :

$$\boldsymbol{f(x,y) = \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} \alpha(u)\alpha(v) C(u,v) cos\left[\frac{\pi(2x+1)u}{2N}\right] cos\left[\frac{\pi(2y+1)v}{2N}\right]}, \text{(III. 8)}$$

Avec $x, y = 0,1,2, \dots, N-1$.

La figure III.7 donne un exemple de la transformée en cosinus appliquée sur une image en niveau de gris.

Figure III.7 : Exemple de la transformée DCT.

III.5. Description de la méthode proposée

Inspirer des caractéristiques de la transformée en cosinus discrète (DCT), elle permet de concentrer l'énergie d'une image dans quelques coefficients, nous proposons une nouvelle méthode de reconnaissance d'activités humaines basée sur les descripteurs globaux.

Nous présenterons deux variantes de la méthode proposée, la première a pour but la reconnaissance des activités dans les séquences vidéo, en utilisant un descripteur calculé par la DCT des cartes spatio-temporelles issues des silhouettes, et les machines à vecteurs support SVM pour la classification. La deuxième variante, est proposée pour la reconnaissance d'activités image par image en temps réel en utilisant un descripteur calculé par la DCT des silhouettes, et les réseaux de neurones artificiels multicouches RBF pour la classification.

a- Descripteur basé sur les cartes spatio-temporelles et la DCT

Dans un but de proposer une méthode simple de reconnaissance d'activités humaines dans les séquences vidéo, nous présentons une nouvelle technique basée sur des cartes spatio-temporelles calculées à partir des squelettes tirés des silhouettes, et la transformée en cosinus discrète DCT pour la création des vecteurs caractéristiques.

L'organigramme de la méthode proposée est présenté dans la figure III.8:

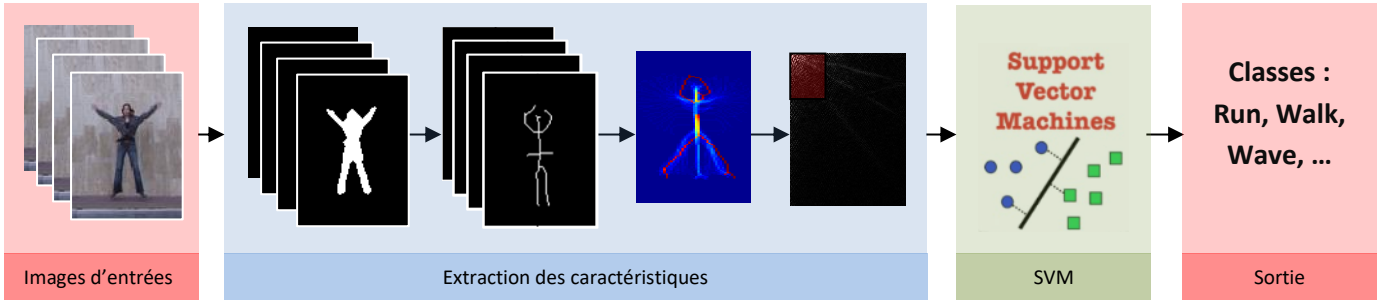


Figure III.8 : Organigramme de la méthode proposée en utilisant les cartes spatio-temporelles.

La première étape de notre méthode est l'extraction des silhouettes à partir des images de la séquence vidéo, pour cela, nous proposons l'utilisation de l'algorithme d'Otsu [81] :

Soit $g(x,y)$ l'image qui contient la silhouette de l'image originale $f(x,y)$, généréeen utilisant le seuil T comme suit [81] :

$$g(x,y) = \begin{cases} 1 & if\ f(x,y) \geq T \\ 0 & otherwise \end{cases} \qquad \text{(III. 9)}$$

La figure III.9 montre des échantillons d'images tirées de la base de données Weizmann et leurs silhouettes extraites par la méthode d'Otsu.

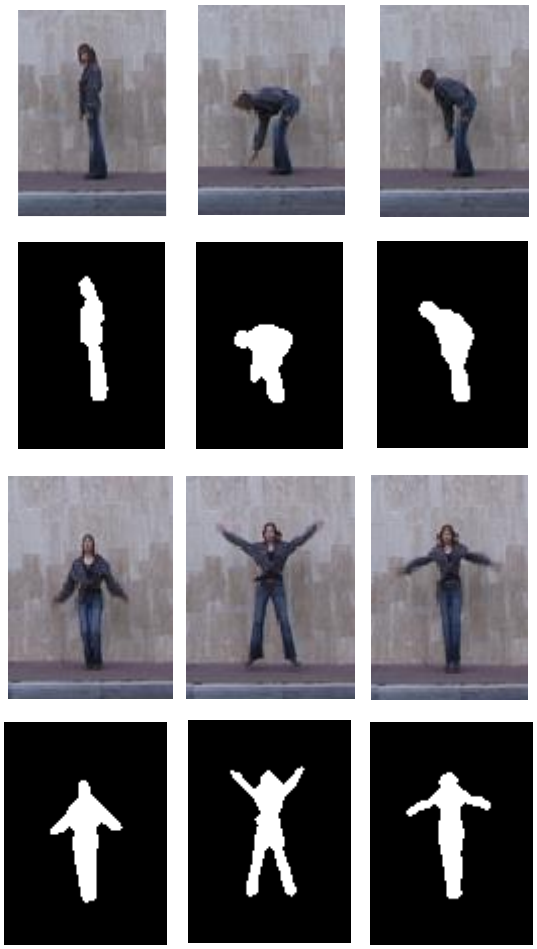


Figure III.9 : Echantillons d'images et leurs silhouettes tirées de la base de données de Weizmann.

Le choix de la méthode d'Otsu est lié à la nature des données disponibles dans les bases de données utilisées dans le domaine, ainsi qu'à la simplicité, l'efficacité et la rapidité de cette technique par rapport aux autres techniques plus complexes.

L'étape suivante est l'extraction des squelettes à partir des silhouettes, nous avons utilisé la méthode *Morphological skeleton* proposée dans [82].

Nous proposons un nouveau descripteur spatio-temporel basé sur les squelettes, ce descripteur est une carte qu'est la combinaison de l'information temporelle est l'information

spatiale des images consécutives de la séquence vidéo, ce descripteur est calculé par l'équation suivante :

$$Features_{map(t+1)} = features_{map(t)} + abs\big(skeleton(t+1) - skeleton(t)\big) \quad \text{(III. 10)}$$

Les cartes spatio-temporelles tirées de la base de données de Weizmann dans la figure III.10 montrent que ce descripteur permet la discrimination entre les activités.

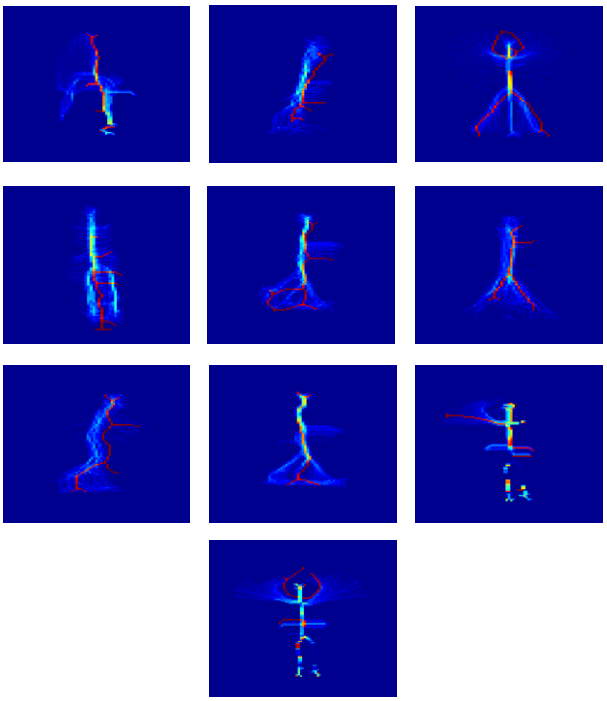


Figure III.10 : Cartes spatio-temporelles calculées sur la base de données de Weizmann.

Pour la création des vecteurs caractéristiques finaux, nous proposons l'utilisation de la transformée en cosinus discrète (DCT). Ce choix est lié au pouvoir de concentration de la DCT de l'information contenue dans les cartes spatio-temporelles en peu de coefficients de taille MxN.

La deuxième étape est l'opération de classification et reconnaissance, nous proposons l'utilisation des machines à vecteurs support SVM avec un noyau non linéaire gaussien (*Support Vector Machines)* multi-classes en utilisant le protocole *« one versus all »*, ces derniers ont montré leur efficacité dans le domaine de la classification. Le modèle SVM final optimisé après apprentissage et validation est utilisé directement pour la reconnaissance des activités dans les séquences vidéo.

b- Descripteur basé sur les silhouettes et la DCT

Les cartes de descripteur présentées précédemment ne permettent pas une reconnaissance en temps réel image par image, parce qu'elles sont basées sur la combinaison de plusieurs images consécutives dans la même carte spatio-temporelle. Afin de proposer un descripteur qui permet la reconnaissance des activités en temps réel, nous présentons une deuxième variante plus simple, basée sur les silhouettes et la transformée en cosinus discrète.

La figure suivante III.11 montre l'organigramme de la technique proposée.

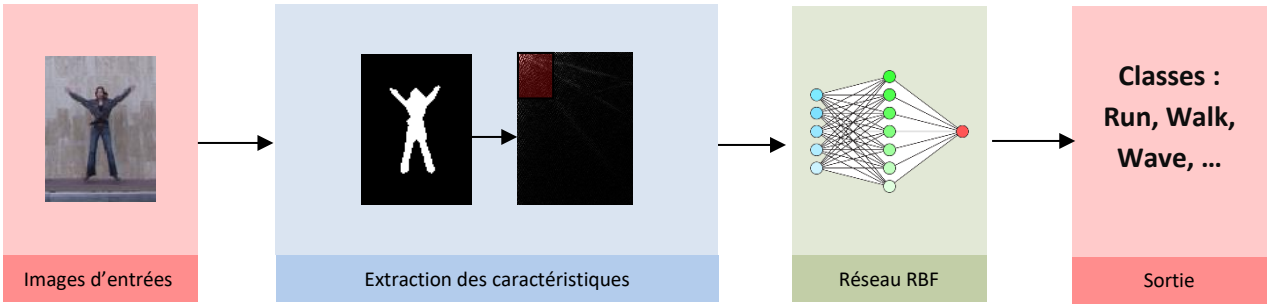


Figure III.11 : Organigramme de la méthode proposée en utilisant les silhouettes.

Pour chaque image du flux vidéo, les silhouettes sont extraites par la méthode d'Otsu, ensuite, on applique la transformée en cosinus discrète directement sur les silhouettes.

Pour la classification, nous proposons l'utilisation des réseaux de neurones artificiels RBF au lieu des machines à vecteurs support (SVM), cela est dû aux limitations de la SVM : quand les données sont très larges, le SVM à tendance a donné des taux de classification moins performants.

III.6. Résultats expérimentaux

Pour l'évaluation des performances de la méthode proposée, nous avons utilisé l'environnement MATLAB pour la création et l'exécution du notre code sur un laptop Core i7 et 8GB de Ram, nous avons réalisé plusieurs tests en utilisant la base de données de Weizmann [5], et avons mené une étude comparative avec les techniques de l'état de l'art.

a. Descripteur basé sur les cartes spatio-temporelles et la DCT

a.1. Protocole de test

Pour l'évaluation des performances de notre approche, nous avons réalisé plusieurs tests en utilisant la base de données de Weizmann. Nous avons mis en place un protocole qui définit plusieurs ensembles de test à partir de la base de données globale, dont le but est de trouver le meilleur partage, qui donne le modèle final le mieux optimisé. Nous n'avons pas utilisé un sous-ensemble de validation à cause de la taille de la base de données. Le tableau III.1 représente le partage de l'ensemble de test :

	Nombre de vidéo dans l'ensemble de test/ nombre total des vidéos pour chaque activité
Test1	3/9
Test2	4/9
Test3	5/9

Tableau III.1 : Protocole de test utilisé dans le cas des descripteurs spatio-temporelles.

Ainsi, afin de réaliser une étude comparative avec les techniques proposées dans la littérature, nous avons réalisé des tests en considérant soit les 10 activités de la base de données de Weizmann soit seulement 9 activées (sans l'activité Skip).

a.2. Discussion des résultats

Le tableau III.2 montre les résultats de reconnaissance obtenues :

	Taux de reconnaissance (%)	
Nombre de vidéo dans l'ensemble de test/ nombre total des vidéos pour chaque activité	***10 activités***	***9 activités***
3/9	90% (27/30)	**100% (0/27)**
4/9	**92.5000% (37/40)**	97.2222% (1/36)
5/9	90% (45/50)	95.5556% (2/45)

Tableau III.2 : Taux de reconnaissance lors de l'utilisation des cartes de descripteurs spatio-temporelles.

Les résultats du tableau III.2 montrent que le meilleur taux de reconnaissance obtenu par la méthode proposée en utilisant la base de données globale (10 activités) est de **92.5%,** par contre, notre méthode a donné un taux de **100%** lors de l'exclusion de l'activité Skip de la base de données. Cela montre que les cartes spatio-temporelles générées ont permis une bonne

séparation des activités, cependant, pour les deux activités Skip et Side, nous avons remarqué une grande ressemblance, ce qui explique les résultats obtenus en utilisant 10 activités dans la base de données.

Le tableau III.3 suivant montre les taux de reconnaissance obtenus dans la littérature :

Méthodes	Taux de reconnaissance
Boiman and Irani 2006 [85]	97.5% (9 activités)
Scovanner et al. 2007 [29]	82.6% (10 activités)
Wang and Suter 2007 [86]	97.8% (10 activités)
Kellokumpu et al 2008 [87]	97.8% (10 activités)
Kellokumpu et al. 2009 [37]	98.7% (9 activités)
Hafiz Imtiaz et al. 2015 [44]	100% (10 activités)
Tasweer et al. 2015 [77]	92.25% (10 activités)
Méthode proposée (DCT+SVM)	**92.5% (10 activités)**
Méthode proposée (DCT+SVM)	**100% (9 activités)**

Tableau III.3 : Comparaison des résultats obtenues par rapport aux autres techniques de l'état de l'art.

L'étude comparative des résultats obtenus par notre méthode et les résultats de l'état de l'art montre que la méthode proposée a donné des résultats comparables en utilisant la base de données globale avec un taux de **92.5%.** Ce taux de reconnaissance atteint les **100%** si l'activité Skip est exclue. Cela est interprété dans la matrice de confusion de la figure III.12, qui montre que la majorité des erreurs de classification sont liées aux activités *Skip* et *Side* à cause de la grande ressemblance entre les deux activités.

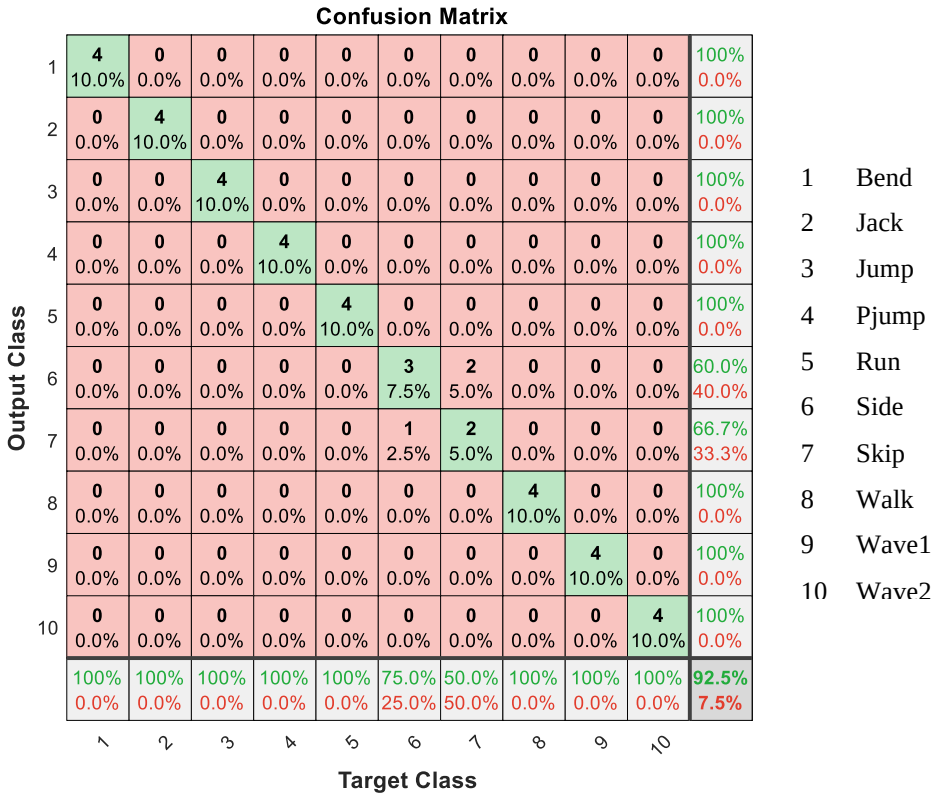


Figure III.12 : Matrice de confusion en utilisant 10 activités.

Lors de l'utilisation de 9 activités seulement de la base de données de Weizmann, la matrice de confusion de la figure III.13 confirme le taux de reconnaissance obtenu de 100%.

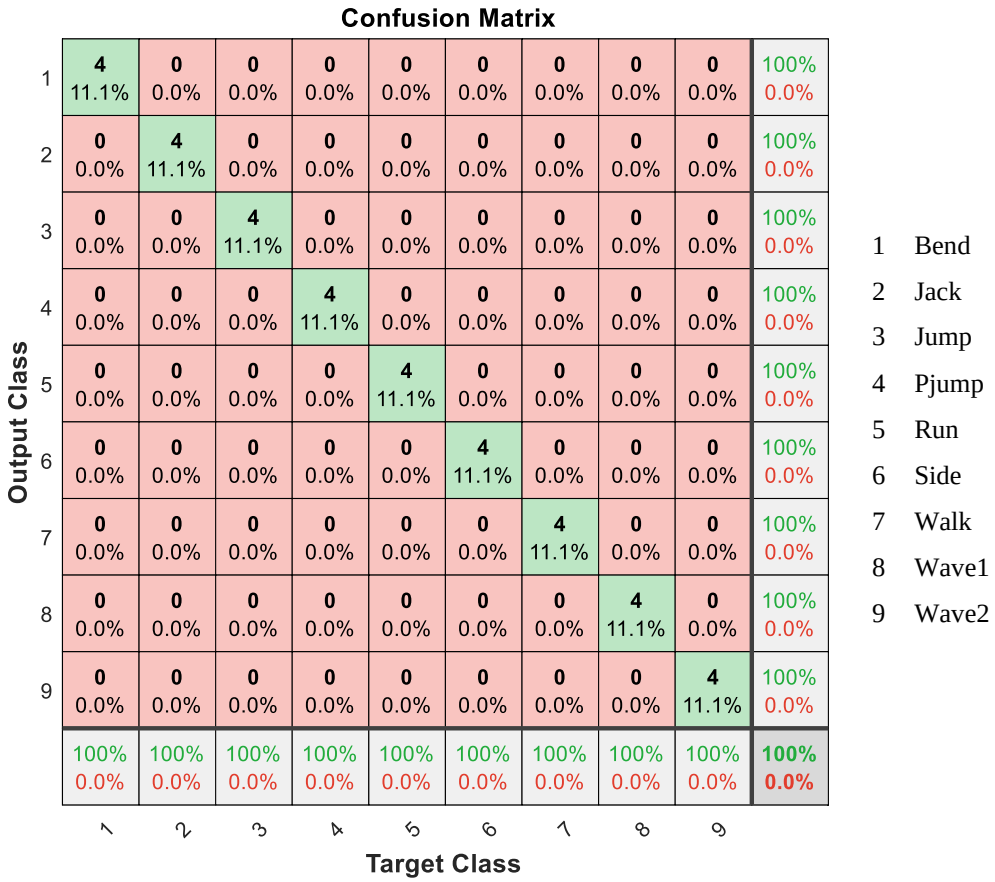


Figure III.13 : Matrice de confusion en utilisant 9 activités.

Les résultats expérimentaux obtenus montrent que la méthode proposée a donné des résultats comparables aux techniques de l'état de l'art, cependant, notre approche est plus simple et plus rapide puisque d'un côté, la reconnaissance est réalisée sur un vecteur caractéristique de taille 8x8 et non pas sur des images de grande dimension, et de l'autre côté, les algorithmes d'extraction des skeletons et des silhouettes utilisées sont simples et rapides.

Le descripteur spatio-temporel proposé dans cette section souffre d'un inconvenant, il est calculé en utilisant les images consécutives de la séquence vidéo, cela rend son utilisation pour la reconnaissance en temps réel impossible, pour remédier à ce problème nous proposons une deuxième approche basée directement sur les silhouettes.

b. Descripteur basé sur les silhouettes et la DCT

b.1. Protocole de test

Pour analyser les performances de cette approche, nous avons utilisé aussi la base de données de Weizmann. Dans ce cas, nous avons divisé la base de données image par image en trois sous-ensembles : entrainement, validation et test comme indiqué dans le tableau III.4 :

	Nombre d'images dans la base de données	entrainment	validation	Test
RBF	100%	40%	30%	30%
	5528	2212	1658	1658

Tableau III.4 : Découpage de la base de données en sous-ensembles pour l'entrainement, la validation et le test

Puisque ici, nous travaillons sur des images tirées des vidéos, la taille de la base de données est considérable, donc on a préféré utiliser le découpage standard du domaine de classification par réseaux de neurones, 40% pour l'ensemble d'entrainement, 30% pour la validation, et 30% pour le test.

b.2. Discussion des résultats

Nous avons effectué plusieurs tests afin de sélectionner les paramètres optimaux de notre technique, à savoir, la dimension du vecteur caractéristiques, le nombre de neurones et la valeur Sigma du réseau de neurones RBF.

À cause que le nombre d'images dans la base de données est très grand, les vecteurs caractéristiques ont une dimension de 64, et dans un but de vérifier si les descripteurs utilisés a permis une bonne séparation des activités, nous avons utilisé le PCA (Analyse en composantes principales) pour la visualisation de la base de données dans un espace en 3D. La figure III.14 présente la projection de la base de données (matrice des caractéristiques) en utilisant la PCA.

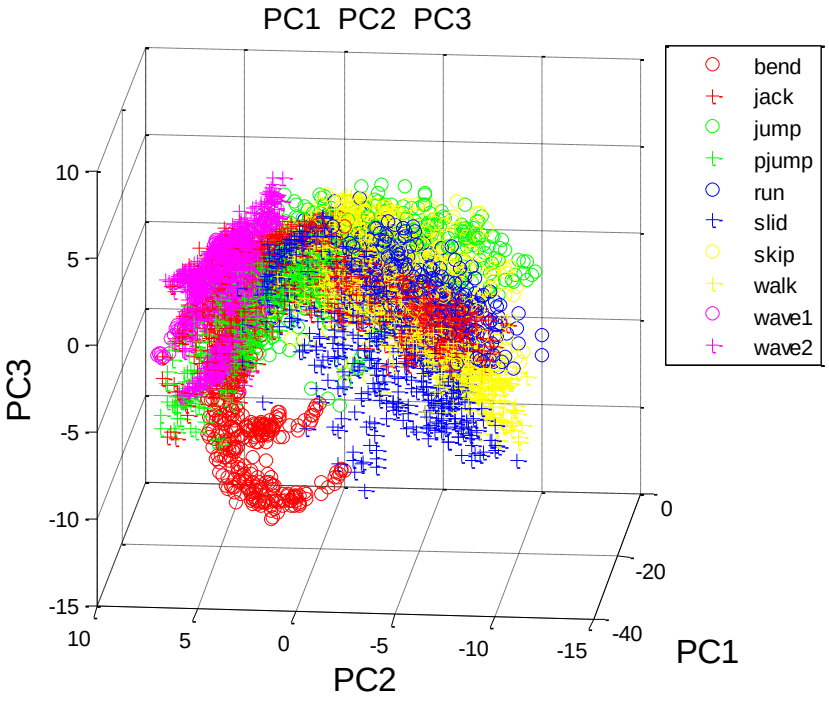


Figure III.14 : Projections des vecteurs caractéristiques tirés de la base de données de Weizmann en utilisant PCA.

La figure III.14 démontre que le descripteur basé sur les silhouettes et la transformée en cosinus discrète (DCT) a permis une bonne séparation des activités.

Pour la sélection de la dimension optimale du vecteur caractéristiques, nous avons réalisé plusieurs tests en utilisant différentes tailles de la fenêtre MxN des coefficients de la transformée en DCT. Les résultats expérimentaux ont montré que la taille de la fenêtre des descripteurs MxN à un impact majeur sur les performances de la technique, le tableau III.5 montre les taux de reconnaissance du test en fonction de la taille de la fenêtre de descripteur DCT.

Taille de la fenêtre des descripteurs	Taux de reconnaissance
3x3	71.4113 %
4x4	83.2931 %
5x5	92.1592 %
6x6	94.0290 %
7x7	97.8287 %
8x8	**99.0350 %**

Tableau III.5 : Taux de reconnaissance en fonction de la taille de la fenêtre des descripteurs.

Les résultats du tableau III.5 montrent que le meilleur taux de reconnaissance a été obtenu lors de l'utilisation d'une fenêtre de descripteurs DCT de taille **8x8,** donc un vecteur caractéristique de taille 64 éléments.

Nous avons réalisé aussi plusieurs tests pour la sélection des paramètres optimaux du modèle RBF final, la figure suivante III.15 montre la variation de taux de validation en fonction du nombre de neurones dans le réseau RBF.

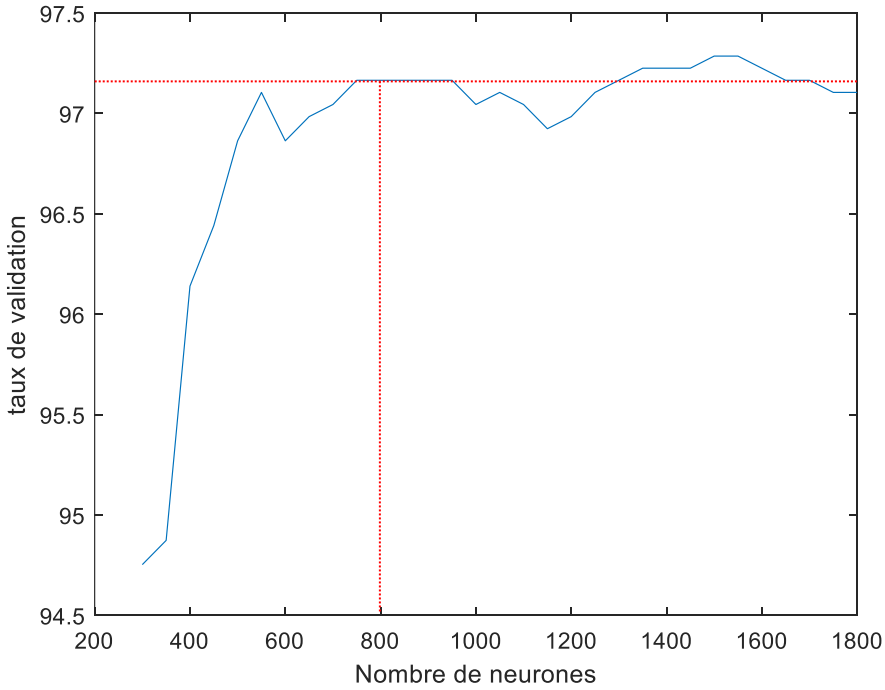


Figure III.15 : Variation du temps de validation en fonction du nombre de neurones.

Les résultats expérimentaux de la figure III.15, ont montré qu'au-delà de **800 neurones** dans le réseau RBF, il n'y a pas une grande différence dans les performances du modèle, cependant, on a remarqué que le modèle devient très lourd et la phase d'apprentissage prend un temps très grand lorsqu'on augmente le nombre de neurones du modèle RBF, par exemple, l'apprentissage du réseau RBF avec **800 neurones** dans la couche cachée se fait dans plus de **6 heures**, pour un réseau avec **1200 neurones**, la procédure prendre plus de **16 heures**, et pour un nombre de neurones de **1600**, l'apprentissage prend plus de **40 heures**.

Dans le but d'optimiser le modèle RBF utilisé, nous avons réalisé plusieurs tests sur la valeur de Sigma. Pendant les tests, nous avons fixé la taille de la fenêtre à 8x8 et le nombre de neurones à 800. Les résultats sont représentés dans la figure III.16.

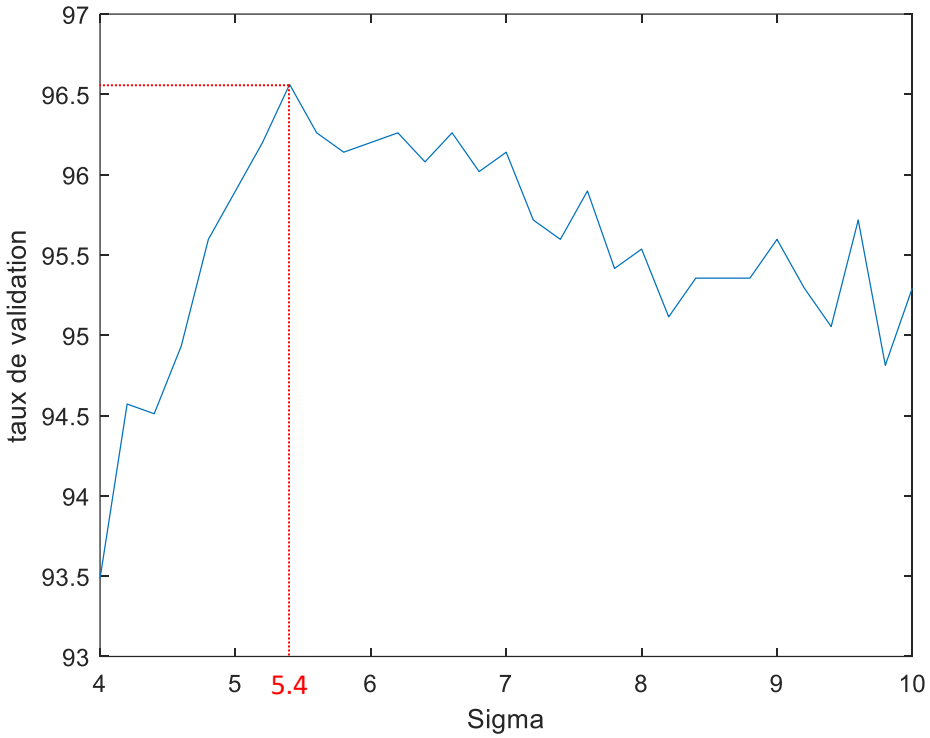


Figure III.16 : Variation du taux de validation en fonction de sigma.

La figure III.16 montre que le meilleur modèle a été obtenu pour une valeur de sigma égale à **5.4**.

Le tableau III.6, montre les paramètres optimaux et les performances du modèle RBF final :

Nombre d'activités	Sigma	Nombre de neurones	Mal classées	Bon classées	Taux de classification (test)
1658	5.4	800	16	1642	**99.0350 %**

Tableau III.6 : les paramètres optimaux du modèle final.

La figure III.17 suivante montre la courbe d'apprentissage du modèle en utilisant les paramètres optimaux obtenues après les tests :

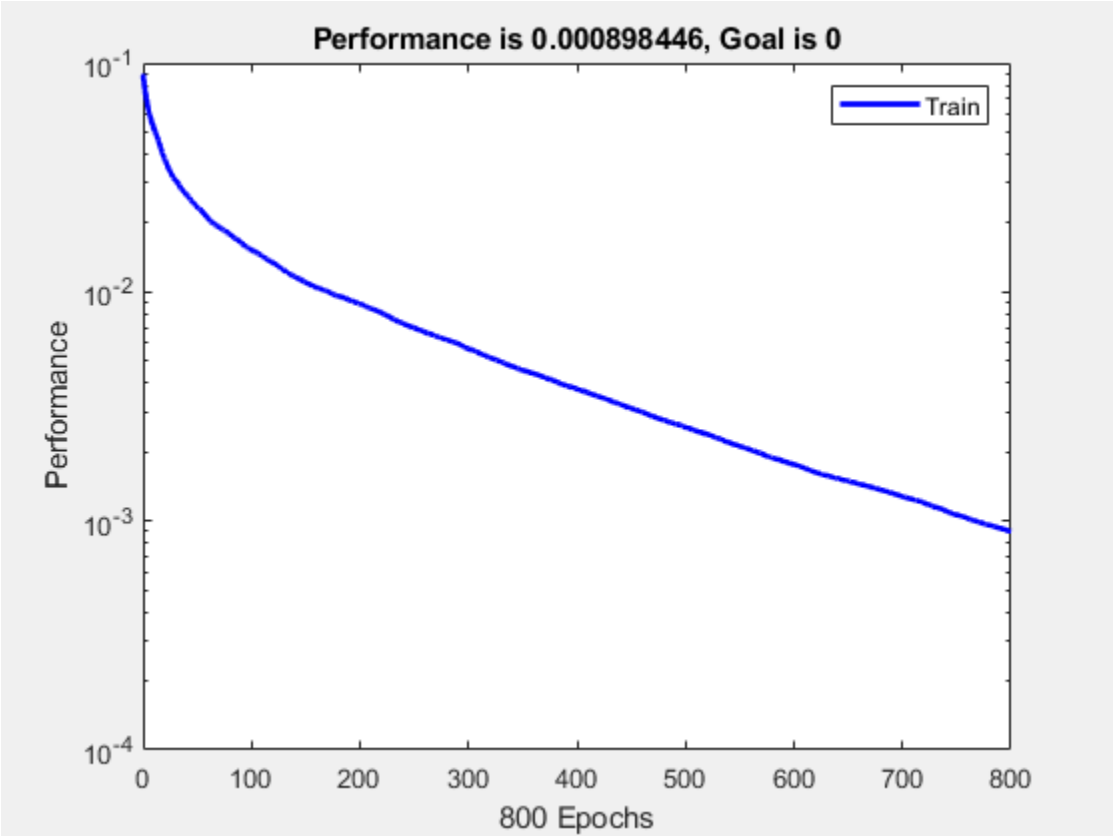


Figure III.17 : Courbe d'apprentissage du modèle RBF.

Les résultats expérimentaux en utilisant le modèle RBF optimisé ont montré que la méthode proposée a permis un taux de reconnaissance de **99.0350%** lors de la reconnaissance image par image. Ainsi, la matrice de confusion de la figure III.18 montre que la méthode proposée a permet d'obtenir un taux de reconnaissance de 100% pour la plupart des activités à l'exception de : run, skip, jump et wave1.

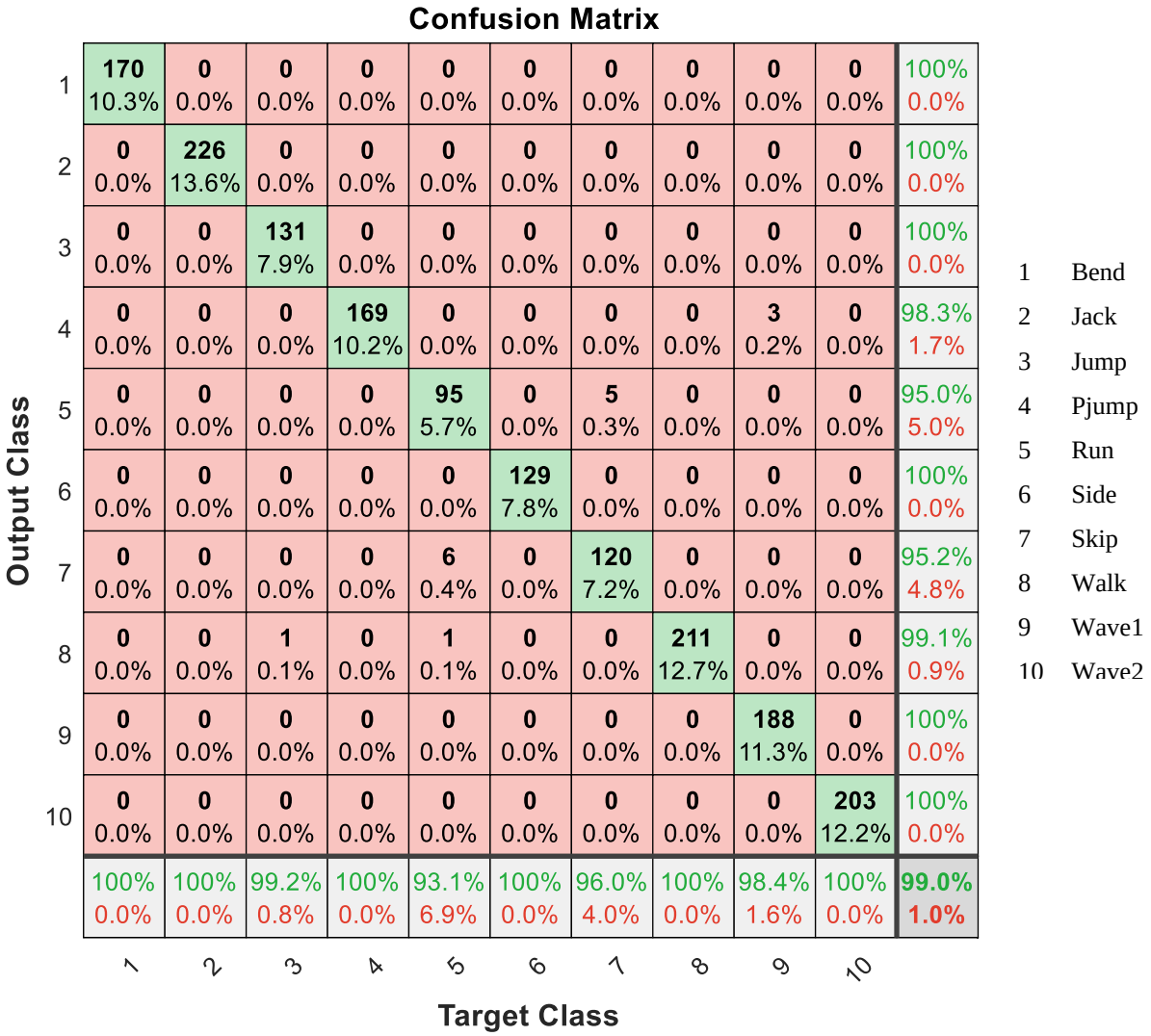


Figure III.18 : Matrice de confusion en utilisant une reconnaissance image par image.

III.8. Conclusion

Dans ce chapitre, nous avons présenté deux variantes de la méthode proposée de reconnaissance d'activités humaines en utilisant la transformée en cosinus discrète DCT. La première variante est dédiée pour la reconnaissance des activités dans les séquences vidéo, elle est basée sur le calcul de la DCT des cartes spatio-temporelles issues par la combinaison des skeletons, el le SVM pour la classification. La deuxième variante a pour but la reconnaissance des activités image par image en temps réel en utilisant la DCT des silhouettes, et les réseaux de neurones artificiels multicouches RBF.

Les résultats expérimentaux ont montré que la méthode proposée pour la reconnaissance des activités dans les séquences vidéo, donne de très bons résultats avec un taux de reconnaissance de **92.5%** et surpasse la plupart des techniques conventionnelles de reconnaissance d'activités humaines en utilisant la base de données de Weizmann.

De même, les résultats obtenus par la deuxième variante pour la reconnaissance des activités image par image sont très performants avec un taux de reconnaissance de **99%** sur la base de données de Weizmann. Cette méthode est efficace, simple à implémenter, rapide et peut être utilisée pour les applications en temps réel.

Les résultats ont montré aussi que la méthode proposée souffre de deux points faibles majeurs, le premier point est le temps d'apprentissage des deux classifieurs utilisés (RBF et SVM), selon la largeur la base de données, cela peut prendre des semaines (une semaine dans notre simulation), le deuxième point est que ces deux classifieurs sont limités lors d'applications sur des problèmes plus complexes tels que la base de données KTH, leurs performances en tendance à baisser, en plus du risque de sur-apprentissage (*Over-Fitting*) et des minimas locaux non-optimaux.

Pour remédier aux limitations de cette méthode nous allons présenter dans le chapitre suivant une nouvelle approche de reconnaissance d'activités humaines en utilisant l'apprentissage profond (*deep learning*).

Partie 2
« Contributions »

Chapitre IV

Reconnaissance d'Activités Humaines en utilisant le Descripteur BSTM et l'Apprentissage Profond

« Maintenant Nous sommes tous connectés par Internet, comme des neurones dans un cerveau géant. »

Stephen Hawking
Artiste, Astronome, Astrophysicien, Cosmologiste, écrivain, écrivain scientifique, Physicien, Scientifique (1942 - 2018)

Chapitre IV Reconnaissance d'Activités Humaines en utilisant le Descripteur BSTM et l'Apprentissage Profond

IV.1 Introduction

Dans le chapitre précédent, nous avons présenté une nouvelle méthode de reconnaissance d'activités humaines en utilisant la transformée en DCT, nous avons présenté deux variantes, une pour la reconnaissance des activités dans les séquences vidéo, et la deuxième pour la reconnaissance des activités image par image en temps réel, cette méthode est idéale pour les problèmes simples avec un nombre de classes très limitées sur des petites bases de données.

Les résultats expérimentaux du chapitre précèdent ont montré que les réseaux de neurones multicouches RBF ont tendance à converger vers des minimas locaux lorsque le nombre de neurones augmente. De plus, le nombre de neurones dans les couches cachées sont très limités, cela rend complexe et difficile leur application dans des problèmes de reconnaissance avec des bases de données très larges telle que KTH.

L'apprentissage profond permet la conception de réseaux de neurones avec plusieurs couches cachées, ce qui lui donne la capacité de résoudre des problèmes beaucoup plus complexes comparé aux réseaux de neurones artificiels multicouches. De plus, l'apprentissage profond permet une auto-extraction des caractéristiques directement à partir des données brutes.

Pour résoudre les problèmes liés à l'utilisation des réseaux de neurones multicouches RBF, et tirer profil des capacités de l'apprentissage profond, nous proposons une nouvelle méthode de reconnaissance d'activités humaines basée sur les réseaux de neurones à convolution CNN, dont l'entrée correspond à un nouveau descripteur spatio-temporel appelé BSTM (*Binary Space-Time Maps*).

IV.2 Description de la méthode proposée

Notre approche est divisée en deux étapes, la première étape consiste en la génération du descripteur BSTM à partir des silhouettes extraites après détection et extraction des zones englobantes des sujets sur un intervalle de temps. Dans la deuxième étape, nous proposons une architecture simple d'un réseau de neurones à convolution CNN pour l'extraction des caractéristiques finales et la reconnaissance.

La figure IV.1 montre l'organigramme de la méthode proposée :

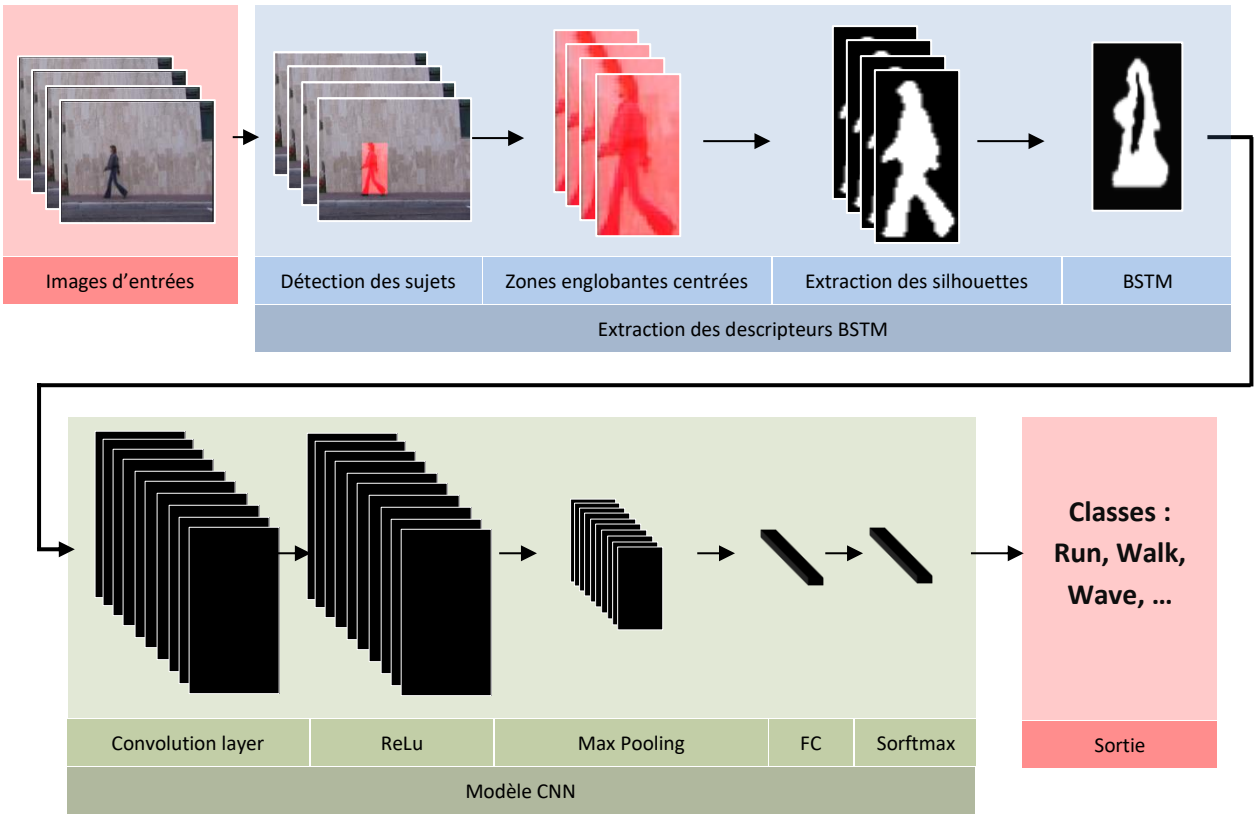


Figure IV.1 : Organigramme de la méthode proposée.

La première opération dans notre méthode est la détection et le suivi des sujets à travers les images de la séquence vidéo, nous avons utilisé un algorithme simple d'extraction des images de fond (figure IV.2), le choix de cet algorithme est lié à la nature des images dans les bases de données utilisées, d'autres types d'algorithmes plus complexes est plus performants peuvent être utilisés selon la nature du problème et selon la complexité des données disponibles.

a) b) c)

Figure IV.2 : Exemple de détection des personnes dans la base de données de Weizmann : a) Walk, b) Side, c) run.

La deuxième opération est l'extraction des zones englobantes des sujets (*Bounding Boxes*), après la détection des sujets dans chaque image, on fait l'extraction du sujet centrée dans une boite englobante de dimension MxN (figure IV.3).

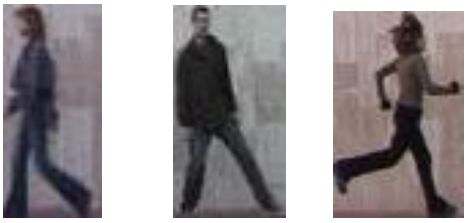


Figure IV.3 : Exemple d'extraction des zones englobantes dans la base de données de Weizmann.

L'opération suivante est l'extraction des silhouettes, pour cela et vu la nature des données nous proposons l'utilisation de l'algorithme d'Otsu [81], d'autres techniques d'extraction de silhouettes peuvent être utilisées. Les silhouettes obtenues par l'algorithme d'Otsu sont des images binaires (figure IV.4).

Figure IV.4 : Exemple d'extraction des silhouettes à partir des zones englobantes dans la base de données de Weizmann.

Après l'extraction des silhouettes, toutes les images sont redimensionnées pour avoir une taille standard permettant la génération de descripteurs BSTM de même taille.

Afin de produire des caractéristiques qui combinent l'information spatio-temporelle des activités, nous proposons un nouveau descripteur spatio-temporel appelé BSTM (*Binary Space-Time Maps*) qui représente la fusion sur un intervalle de temps de l'information temporelle et l'information spatiale des sujets.

Les images BSTM sont calculées à partir des silhouettes obtenues après extraction des zones englobantes des sujets, le tableau IV.1 résume l'algorithme de calcul des descripteurs BSTM :

Variables :

Entrée : Silhouettes $g(x, y)$

Sortie: BSTM (Binary space-time map $h(x, y)$)

begin

For i=1:T

h(x,y)=$\sum abs(g_{ti+1}(x, y) - g_{t1}(x, y))$

End

If h(x,y)≠0

h(x,y)=1

Else

h(x,y)=0

End

End

End

Tableau IV.1 : Algorithme de calcul des descripteurs BSTM.

Avec:

$g(x, y)$: Les images silhouettes,

$h(x, y)$: carte spatio-temporelle binaire (*Binary Space-Time Map*),

T: Le nombre d'images**.**

Les descripteurs BSTM contiennent les informations spatio-temporelles dans un intervalle de temps (T), la qualité des cartes BSTM dépend de la qualité des silhouettes et aussi du nombre d'images utilisées lors de la génération.

Les résultats expérimentaux ont montré que 16 images sont suffisantes pour la génération de bons descripteurs BSTM.

La figure IV.5 montre un exemple de descripteurs BSTM extraits à partir de la base de données « Keck Gesture Database ».

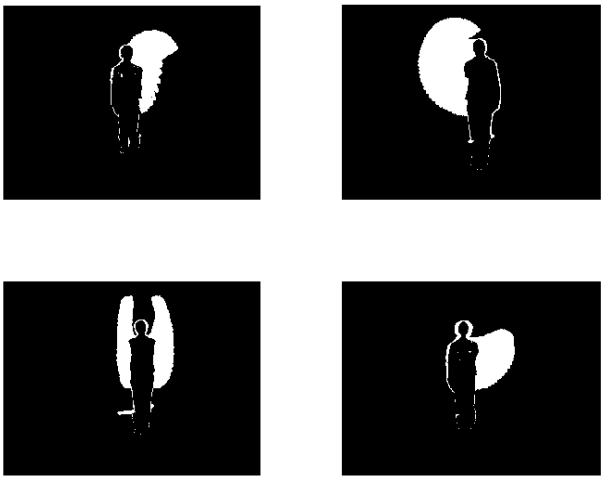


Figure IV.5 : Exemple de descripteurs BSTM en utilisant la base de données « Kech Gesture Database ».

Le fait que les descripteurs BSTM soient calculés à partir des zones englobantes et centrées sur les sujets permet à notre approche de détecter plusieurs personnes en même temps, ce qui rend possible la reconnaissance de plusieurs activités dans la même séquence vidéo.

De plus, notre approche à l'avantage d'être simple, rapide et ne nécessite pas beaucoup de ressources machine, grâce à la nature même des descripteurs BSTM qui sont des images binaires contenant l'information spatio-temporelle de l'activité dans un intervalle de temps réduit.

La deuxième étape dans cette méthode est la génération des descripteurs finaux et la reconnaissance des activités. Pour cela, nous proposons l'utilisation des réseaux de neurones à convolution CNN.

Les entrées de notre réseau ConvNet sont les images BSTM définis sur un intervalle de temps, l'architecture du réseau CNN utilisé est la suivante :

Une couche d'entrée : à la même dimension que les images BSTM.

Une couche de convolution : Les paramètres de cette couche en utilisant les trois bases de données sont présentés dans le tableau IV.2:

Base de données	Taille de filtres de convolution	Nombre de filtres	Stride
Weizmann	[6x6]	16	1
Keck Gesture Dataset	[5x5]	10	1
KTH	[3x3]	10	1

Tableau IV.2 : Paramètres de la couche de convolution.

Un exemple d'activation de la couche de convolution pour l'activité Running sur la base de données de Weizmann est illustré par la figure IV.6:

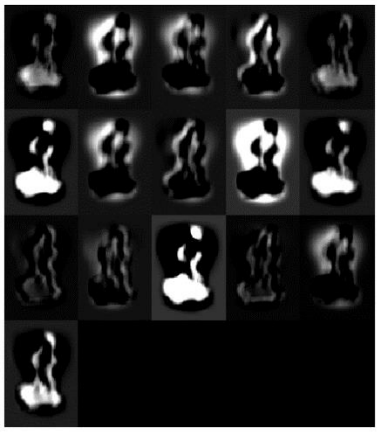
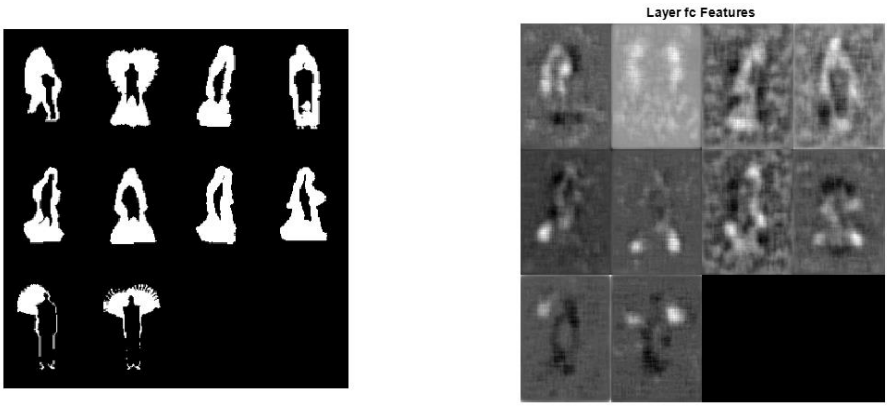


Figure IV.6 : Les 16 activations de la couche de convolution de l'activité « Running » dans la base de données de Weizmann.

Une couche Relu : qui réalise une opération de seuillage, toutes les entrées inférieures à zéro sont remises à zéro.

Une couche Maxpool : cette couche réalise une opération de rééchantillonnage en utilisant la fonction max dans une fenêtre de taille 2x2.

Une couche entièrement connectée : cette couche contient les descripteurs de chaque activité présentée à l'entrée. La figure IV.7 montre un exemple des BSTM et leurs descripteurs issus de la couche entièrement connectée.



a) b)

Figure IV.7 : a) Exemple de BSTM de chaque activités, b) Les descripteurs issus de la couche entièrement connectée (base de données de Weizmann).

Une couche Softmax : la sortie de cette couche est un vecteur de probabilité de chaque classe calculée en utilisant la fonction d'activation sigmoïde.

La couche de classification : cette couche est responsable d'interpréter le vecteur Softmax est de délivrer les classes à la sortie du réseau.

IV.3 Résultats expérimentaux

Pour l'évaluation des performances de la méthode proposée, nous avons réalisé plusieurs tests en utilisant les trois bases de données (Weizmann, Keck Gesture Database et KTH) lors de la simulation. Nous avons utilisé un laptop i7 avec 8GO de Ram et 2 GO de carte graphique NVIDIA.

a. La base de données de Weizmann

La base de données de Weizmann est constituée de 10 activités réalisées par 9 personnes, nous avons adopté le protocole de test suivant : 4 personnes dans chaque activité (40 séquences vidéo) pour l'apprentissage et 5 personnes dans chaque activité (50 séquences vidéo) pour le test.

La figure IV.8 montre un exemple de la base de données de Weizmann, son BSTM et l'activation de la couche de convolution correspondante :

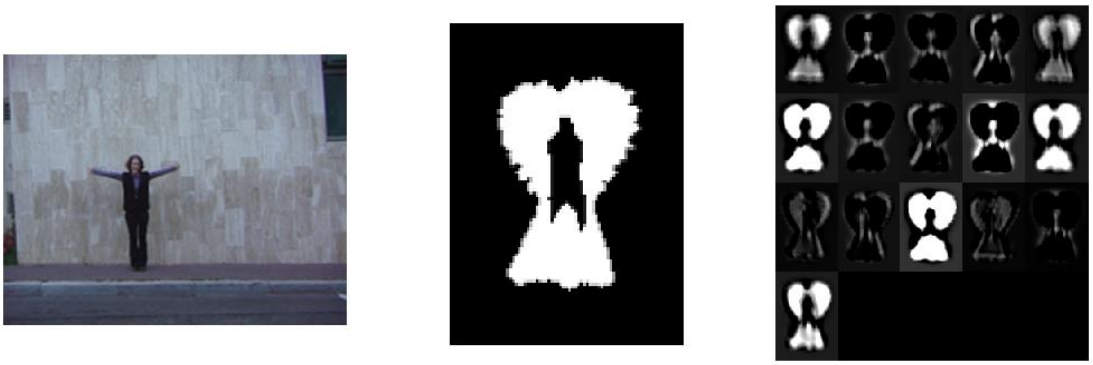


Figure IV.8 : De gauche à droite, échantillons d'image de la base de données de Weizmann, son BSTM et l'activation de la couche à convolution correspondante.

Le tableau IV.3 suivant montre les taux de reconnaissance de la méthode proposée par rapport aux techniques de reconnaissance d'activités humaines conventionnelles ainsi que les techniques basées sur l'apprentissage profond présentées dans l'état de l'art.

Méthode	**Taux de reconnaissance**
Boiman and Irani 2006 [**85**]	97.5% (9 actions)
Scovanner et al. 2007 [**29**]	82.6% (10 actions)
Wang and Suter 2007 [**86**]	97.8% (10 actions)
Kellokumpu et al 2008 [**87**]	97.8% (10 actions)
Kellokumpu et al. 2009 [**37**]	98.7% (9 actions)
Hafiz Imtiaz et al. 2015 [**44**]	100% (10 actions)
Tasweer et al. **2015 [77**]	92.25% (10 actions)
Tushar et al. 2015 [14]	100% (5 actions)
Méthode proposée (DCT+SVM)	**92.50% (10 actions)**
Méthode propose (BSTM+CNN)	**98% (10 actions)**

Tableau IV.3 : Taux de reconnaissance en utilisant la base de données de Weizmann.

L'étude comparative en utilisant les taux de reconnaissance dans le tableau IV.3 montre que la méthode proposée surclasse les autres techniques conventionnelles ainsi que celles basées sur l'apprentissage profond avec un taux de reconnaissance de **98%**.

La comparaison des taux de reconnaissances entre les deux techniques proposées en utilisant 10 activités dans la base de données de Weizmann, montre que la technique basée sur l'apprentissage profond donne un taux de reconnaissance de **98%** et surpasse celle basée sur la DCT et le réseau RBF qui a donné un taux de **92.5%.** Cela confirme d'une part l'efficacité

de notre descripteur proposée (BSTM) et d'autre part par l'utilisation d'un réseau CNN pour la reconnaissance des activités.

La courbe d'apprentissage de la figure IV.9, montre que le modèle a rapidement atteint le maximum du taux d'apprentissage, cela montre que les cartes BSTM proposées, ont permis une très bonne séparation des activités pour la base de données de Weizmann.

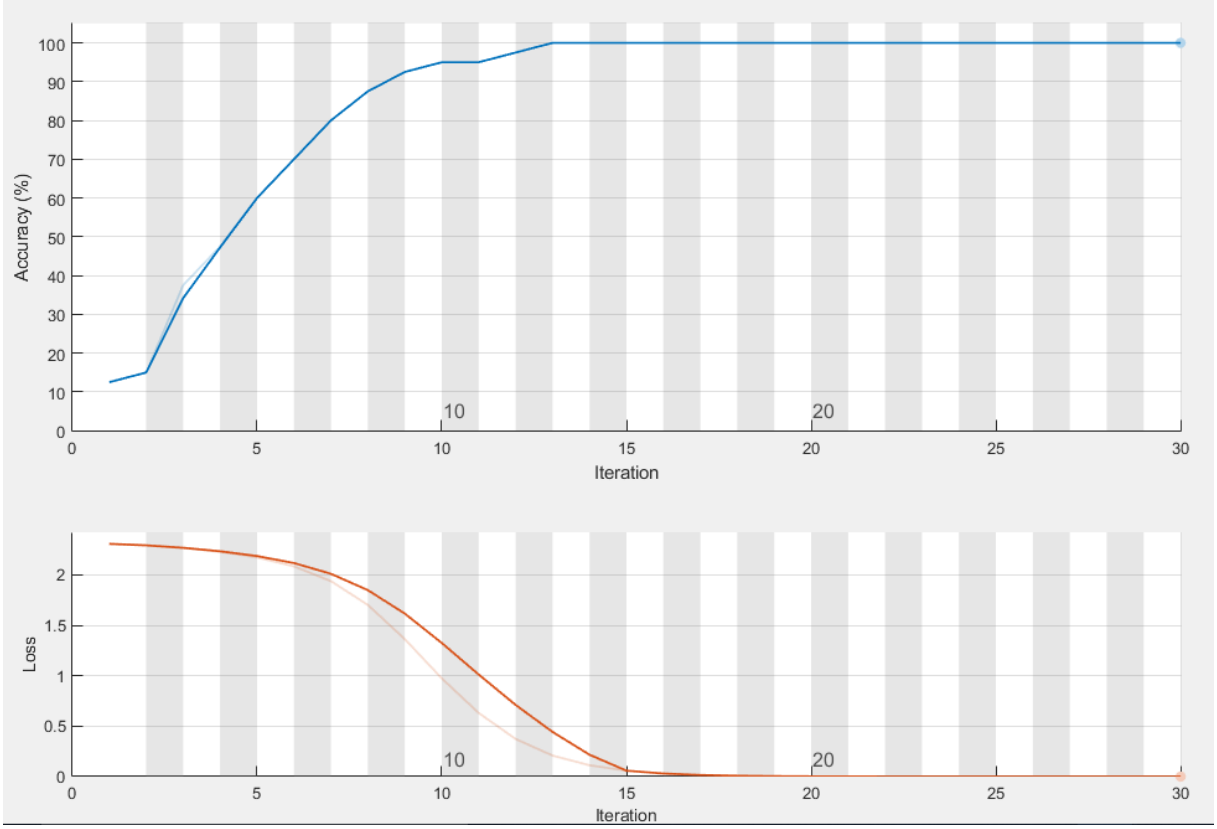


Figure IV.9 : La courbe d'apprentissage en utilisant la base de données de Weizmann.

La matrice de confusion de la figure IV.10 confirme que notre approche permet un taux de reconnaissance de 100% pour toutes les activités dans la base de données de Weizmann, à l'exception de l'activité Run qui a enregistrée une seule fausse classification en Skip. Pour inspecter cette erreur, nous représentons dans la figure IV.11 les descripteurs BSTM des activités Run et Skip. Cette représentation affirme que cette fausse classification est liée à la ressemblance du descripteur BSTM de cette activité et les cartes BSTM de l'activité Skip. Malgré ça, on peut remarquer dans la matrice de confusion que toutes les séquences de l'activité Skip ont été classifier avec un taux de 100%, cela confirme l'efficacité de classification de notre réseau de neurones à convolution CNN.

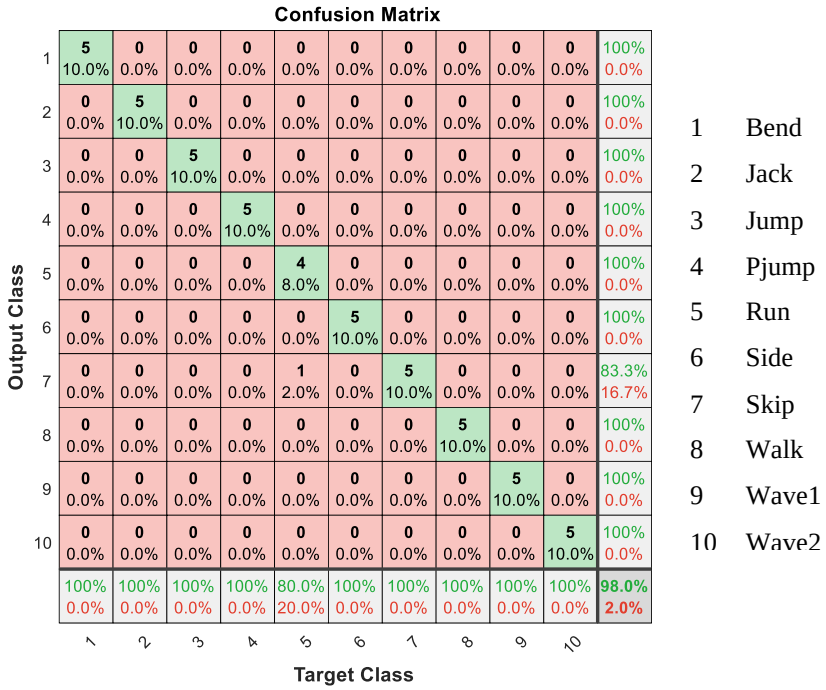


Figure IV.10 : Matrice de confusion en utilisant la base de données de Weizmann.

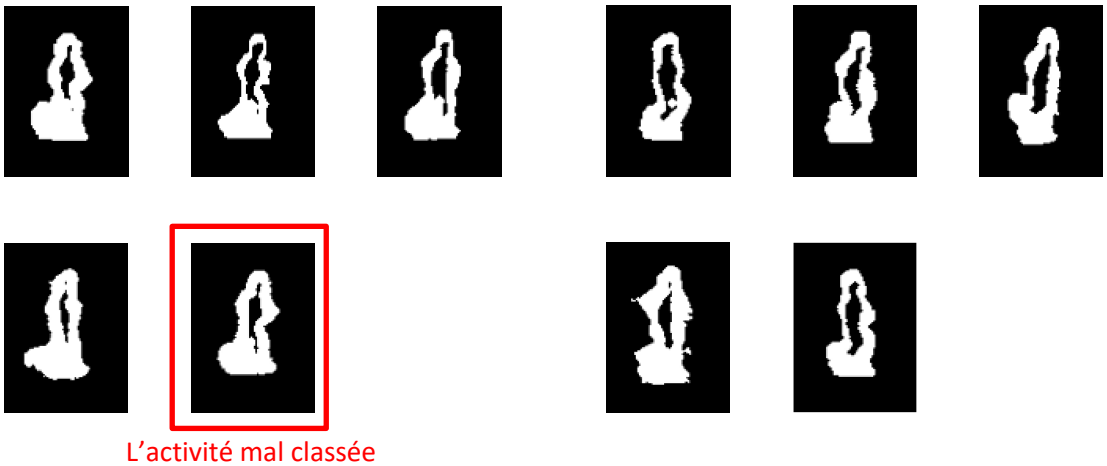


Figure IV.11 : Comparaison des descripteurs BSTM de l'ensemble de test des deux activités : a) Run, b) Skip.

La courbe ROC de la figure IV.12, montre que le modèle final est optimal, avec des courbes ROC parfaites pour toutes les activités et l'aire sous la courbe moyen AUC égale à **1**.

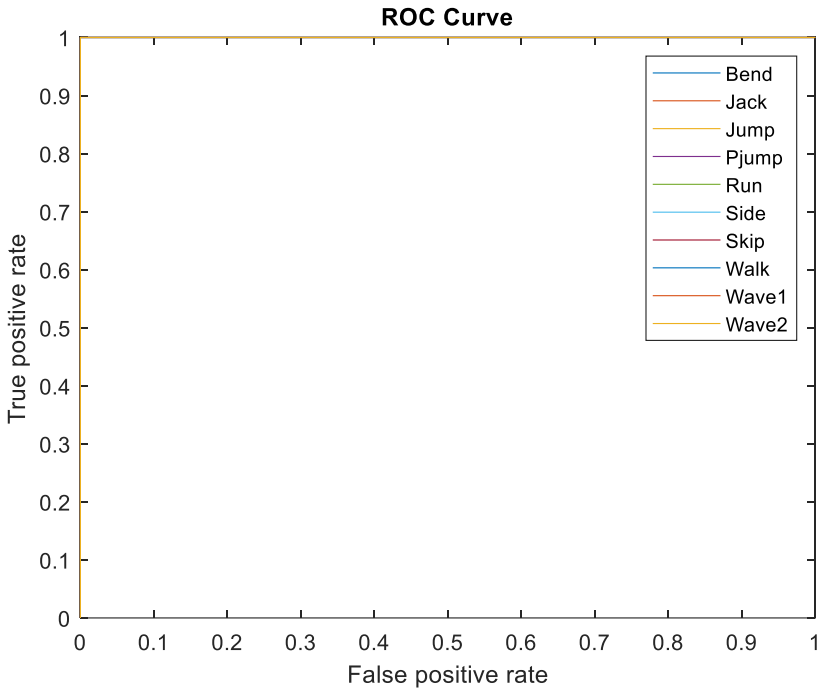


Figure IV.12 : Courbe ROC en utilisant la base de données de Weizmann.

b. Keck Gesture Dataset

La base de données de Keck Gesture Database est constituée de 14 activités réalisées par 3 personnes, elle est composée de 42 séquences vidéo seulement. L'utilisation de cette base constitue un vrai challenge vu sa taille limitée. Dans notre cas, nous avons adopté le protocole de test suivant : utiliser deux personnes de chaque activité pour l'apprentissage et une personne pour le test (28 séquences vidéo pour l'apprentissage et 14 pour le test).

La figure IV.13 montre un exemple d'une activité de la base de données Keck Gesture Dataset, son BSTM et les activations de la couche de convolution.

Figure IV.13 : De gauche à droite, échantillon d'image de l'activité Turn left, son BSTM et les activations de la couche de convolution.

Le tableau IV.4 montre les taux de reconnaissances de la méthode proposée en comparaison avec les résultats obtenus dans la littérature.

Méthode	Taux de reconnaissance
Zhuolin Jiang et al. 2012 [79]	97.5% (9 activités)
Méthode proposée (BSTM+CNN)	**100% (14 activités)**

Tableau IV.4 : Taux de reconnaissance en utilisant la base de données de Keck Gesture Database.

Les résultats du tableau IV.4 montrent que la méthode proposée a donné un taux de reconnaissance de **100%** en utilisant 14 activités et surpasse la méthode de Zhuolin et al. qui a donné un taux de **97.5%** en utilisant seulement 9 activités de la base de donnée.

Quoique que la base de données de Keck Gesture Database soit une petite base de données, La figure IV.14 montre que la méthode proposée a permis aussi une optimisation rapide du modèle CNN proposé.

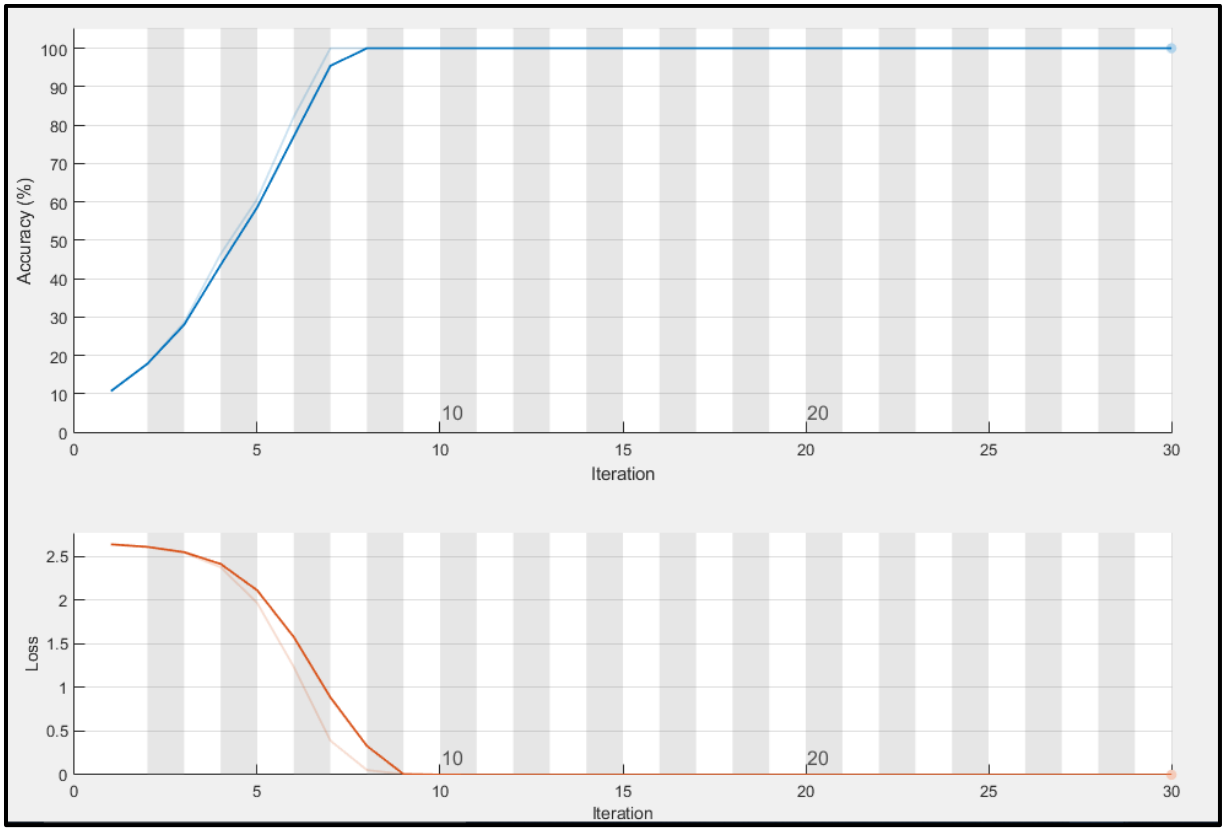


Figure IV.14 : La courbe d'apprentissage en utilisant la base de données de Keck Gesture Database.

La matrice de confusion dans la figure IV.15 montre une reconnaissance parfaite avec des taux de de classification de **100%** pour toutes les activités en utilisant la base de données Keck Gesture Database, ce qui confirme la robustesse du réseau CNN proposé.

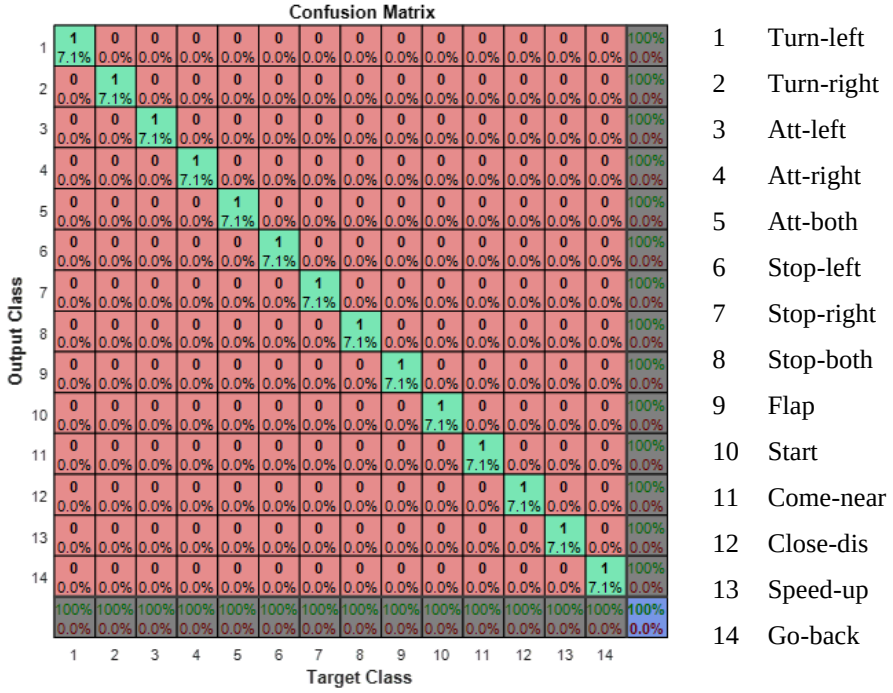


Figure IV.15 : Matrice de confusion en utilisant la base de données de Keck Gesture Database.

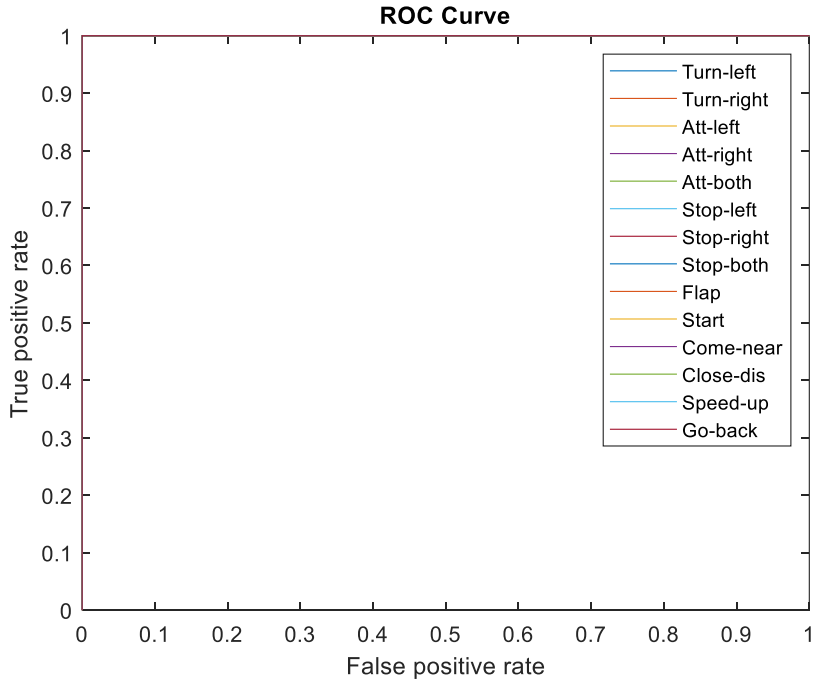


Figure IV.16 : ROC Curve en utilisant la base de données de Keck Gesture Database.

La courbe ROC de la figure IV.16, représente les performances du modèle finale, il est clair que le comportement du modèle est stable pour toutes les activités, avec des courbes Roc idéales, et l'aire sous la courbe moyen AUC égale à **1**.

c. La base de données KTH

La base de données KTH est la base de données la plus utilisée dans le domaine de la reconnaissance d'activités humaines, elle est composée de 6 actions réalisées par 25 personnes dans 4 scénarios différents (600 séquences vidéo). Nous avons adopté le protocole de tests suivants : 480 séquences vidéo pout l'apprentissage et 120 séquences vidéo pour le test.

Un exemple de la base de données KTH, son BSTM et les activations de la couche de convolution sont représentées sur la figure IV.17:

Figure IV.17 : De gauche à droite : échantillon d'image de la base de données KTH, son BSTM et les activations issues de la couche de convolution.

Le tableau IV.5 montre les taux de reconnaissance obtenus en utilisant la base de données KTH :

Méthode	Taux de reconnaissance
Wong and Cipolla. 2007 [88]	86.62%
Niebles et al. 2007 [89]	83.33%
Laptev et al. 2008 [46]	92.10%
Schuldt et al. 2004 [45]	71.70%
Dollar et al. 2005 [27]	81.20%
Bo Chen et al. 2010 [90]	91.13%
Vivek et al. 2015 [91]	93.96%
Lin Sun et al. 2014 [92]	93.10%
Moez B et al. 2015 [15]	94.39%
Méthode proposée (BSTM+CNN)	**92.50%**

Tableau IV.5 : Résultats de reconnaissance en utilisant la base de données de KTH.

L'étude comparative des taux de reconnaissance obtenus lors de l'utilisation de la base de données KTH montre que notre approche donne de très bonnes performances et surpasse la majorité des méthodes conventionnelles de reconnaissance d'activités humaines, elle donne aussi des résultats comparables aux méthodes récentes basées sur l'apprentissage profond.

Les résultats obtenus sont encourageants, parce que contrairement aux autres techniques basées sur l'apprentissage profond, la méthode proposée est plus simple, elle est basée sur le descripteur BSTM qui est simple à extraire à partir des silhouettes des zones englobantes des sujets. Elle utilise un réseau de neurones à convolution CNN très simple composé seulement d'un seul canal est une seule couche de convolution. Ce qui permet de rendre l'approche plus rapide et ne nécessitant pas de grosses ressources matérielles.

Les résultats expérimentaux ont montré que la qualité des silhouettes extraites a un impact direct sur les performances de la méthode proposée, l'utilisation d'algorithmes d'extraction de silhouettes plus performants va sûrement améliorer les performances de cette approche.

La figure IV.18 montre la courbe d'apprentissage en utilisant la base de données KTH :

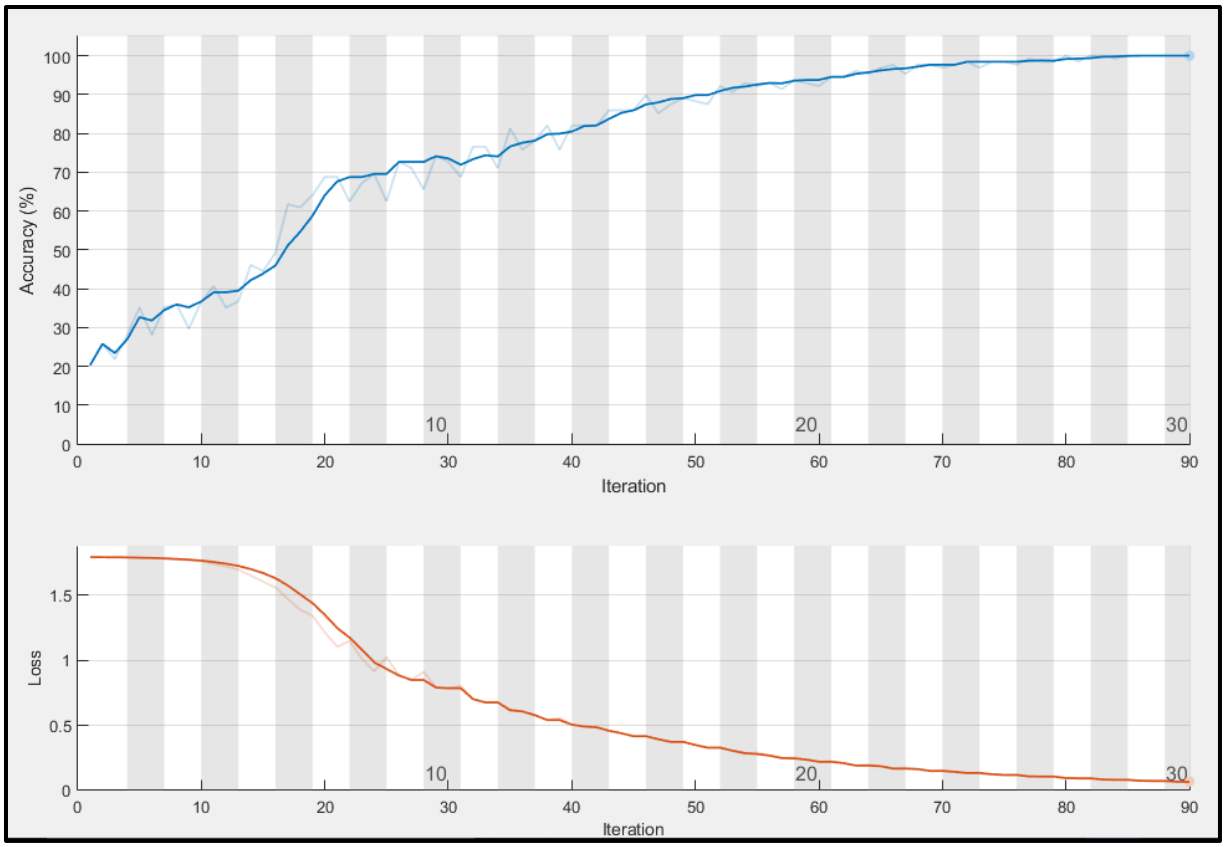


Figure IV.18 : La courbe d'apprentissage en utilisant la base de données KTH.

La matrice de confusion de la figure IV.19, montre que la méthode proposée a permis un taux de reconnaissance de 100% pour les actions : Boxing, Waving et Hand-clapping, et donne un taux de 95% pour l'activité Walking. La majorité des fausses classifications sont liées aux activités Running et Jogging cela est expliqué par la ressemblance entre ces deux activités.

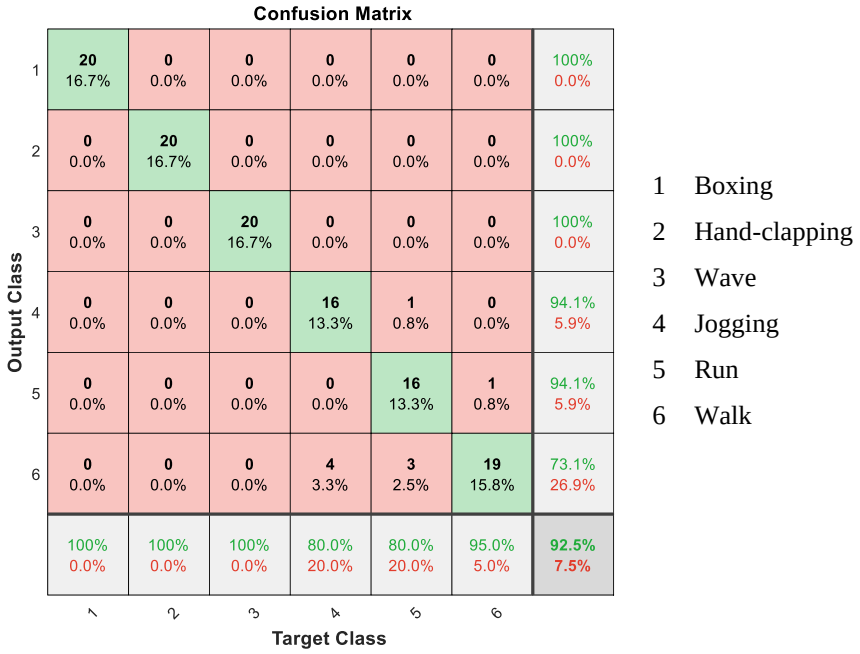


Figure IV.19 : Matrice de confusion en utilisant la base de données KTH.

La courbe ROC de la figure IV.20 montre les performances du modèle CNN final après apprentissage en utilisant la base de données de KTH. Cette courbe montre que le comportement du modèle est stable pour toutes les activités, avec des réponses parfaites pour les activités Boxing, Hand-clapping et Waving, cela est confirmé avec l'aire sous la courbe moyen des activités qui est de **0.9812**

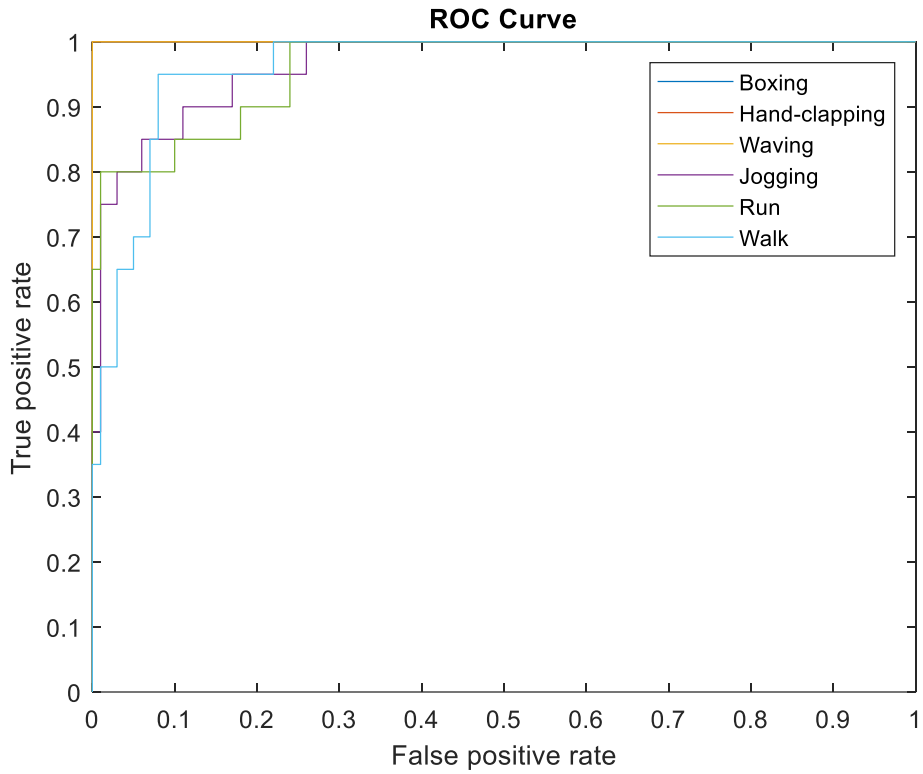


Figure IV.20 : Le ROC Curve en utilisant la base de données KTH.

IV.4. Conclusion

Dans ce chapitre, nous avons présenté une nouvelle méthode de reconnaissance d'activités humaines en utilisant l'apprentissage profond.

Nous avons proposé un nouveau descripteur spatio-temporel BSTM (*Binary Space-Time Maps*) basé sur l'extraction des silhouettes dans les zones englobantes centrées sur les sujets.

Les cartes BSTM ont permis la combinaison de l'information temporelle et l'information spatiale des images de la séquence vidéo dans un intervalle de temps réduit, ils représentent l'évolution spatio-temporelle de l'activité dans ces zones englobantes.

Les résultats expérimentaux ont montré que la méthode proposée est performante sur toutes les bases de données utilisées. Elle est simple, efficace, rapide, peu gourmande en termes de ressources machine est peut-être utilisée pour la détection de plusieurs activités dans la même séquence vidéo grâce à l'utilisation des descripteurs BSTM. Cependant, la méthode proposée est basée sur une étape de prétraitement pour le calcul des cartes BSTM.

En plus, elle ne peut pas être utilisée pour la reconnaissance des activités en temps réel. Dans le chapitre suivant, nous allons proposer une nouvelle approche pour la reconnaissance des activités image par image en temps réel et dans les séquences vidéo totalement automatisée, basée sur l'apprentissage profond et sans aucune étape de prétraitement.

Partie 2
« Contributions »

Chapitre V

Reconnaissance d'Activités Humaines en utilisant le Modèle YOLO

« La réalité virtuelle sera une technologie importante. Je suis assez confiant à ce sujet. »

Mark Zuckerberg
Homme d'affaire, Fondateur de Facebook

V.1. Introduction

Dans le chapitre précédent, nous avons proposé une méthode de reconnaissance d'activités humaines en utilisant un nouveau descripteur spatio-temporel BSTM (*Binary Space-Time Maps*) et l'apprentissage profond. Les résultats expérimentaux ont montré que la méthode proposée est rapide, simple et performante. Cependant, notre méthode est basée sur une étape de prétraitement pour le calcul des descripteurs BSTM dans un intervalle de temps, cela rend son application pour la reconnaissance des activités image par image en temps réel impossible.

Dans ce chapitre, et dans le but de proposer une méthode de reconnaissance d'activités humaines image par image en temps réel, et pour profiter du pouvoir d'auto-extraction des descripteurs de l'apprentissage profond, nous présentons une technique basée directement sur les images brutes en utilisant seulement un réseau de neurones à convolution CNN.

Notre approche est basée sur l'utilisation de l'architecture YOLO (*You Only Look Once*) qui a prouvé son efficacité dans le domaine de la reconnaissance d'objets.

V.2. Présentation du YOLO (*You Only Look Once*)

Le YOLO [83] est un modèle d'apprentissage profond de référence dans le domaine de la reconnaissance d'objets en temps réel basé sur les réseaux de neurones à convolution (CNN). La traduction de YOLO en français c'est « vous regarder seulement une fois » qui est le principe de cette architecture, le même réseau CNN est appliqué directement sur toute l'image une seule fois (ce n'est pas le cas par exemple des R-CNN qui sont basés sur la réalisation de plusieurs passages sur la même image. L'architecture du réseau permet la division de l'image en plusieurs régions et ainsi de prédire les classes et les zones englobantes (*bounding boxes)* pour chaque région. Les régions avec des probabilités élevées sont alors retenues comme une classe (figure V.1). Ce principe rend le modèle YOLO très rapide, 1000 fois plus rapide que les R-CNN et 100 fois que fast R-CNN.

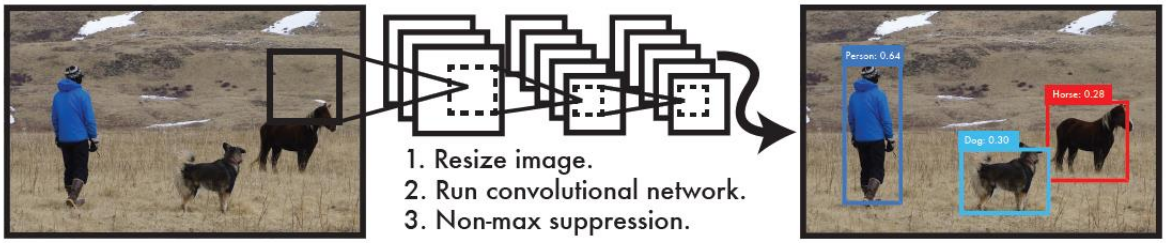


Figure V.1 : Principe du Système YOLO [83].

Les sorties de l'architecture YOLO sont les coordonnées des zones englobantes (*bounding boxes)* et les probabilités des classes de chaque zone, la figure V.2 montre un exemple de reconnaissance d'objets par le modèle YOLO.

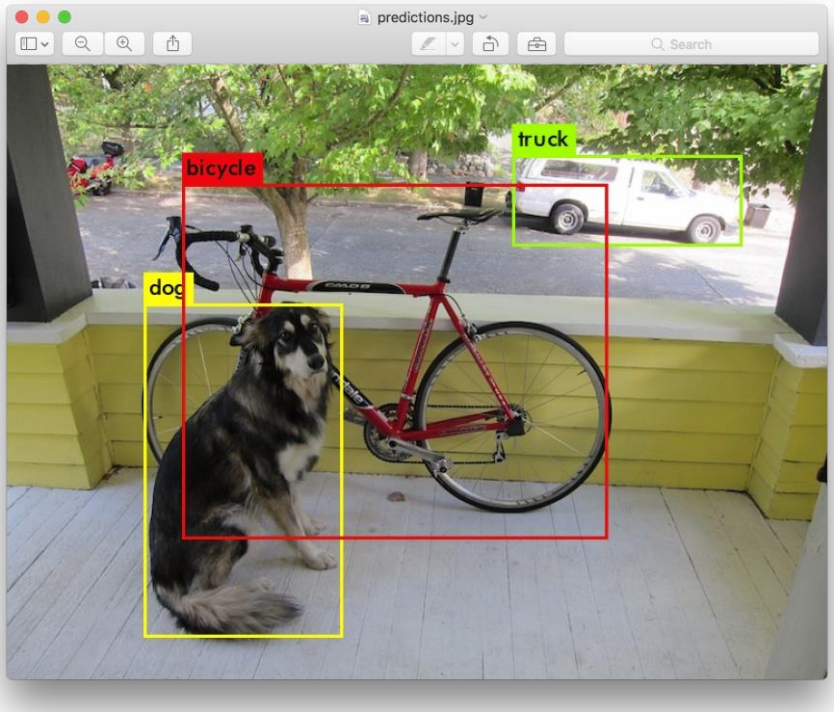


Figure V.2 : Exemple de détection d'objets par YOLO [84].

V.3. Architecture du YOLO

L'architecture du YOLO est composée de 24 couches de convolution (figures V.3 et 5.5) suivi de deux couches entièrement connectées (*fully-connected layers*), la dimension des images d'entrée est de 448x448.

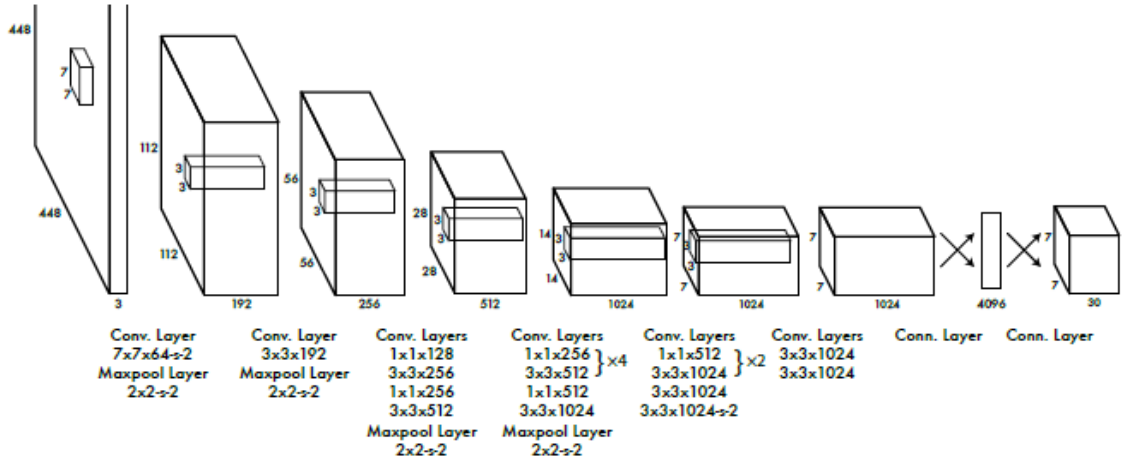


Figure V.3 : Architecture du YOLO [83].

Le principe du modèle est basé sur l'utilisation d'une succession de couches de convolution avec différentes tailles et nombre de filtres, cette architecture permet de diviser l'image en plusieurs zones qui englobent les différents objets ou portions d'objets dans l'image, l'application d'un seuillage des probabilités de chaque classe dans chaque zone permet de conserver uniquement les zones englobantes représentatives qui limitent les objets (figure V.4).

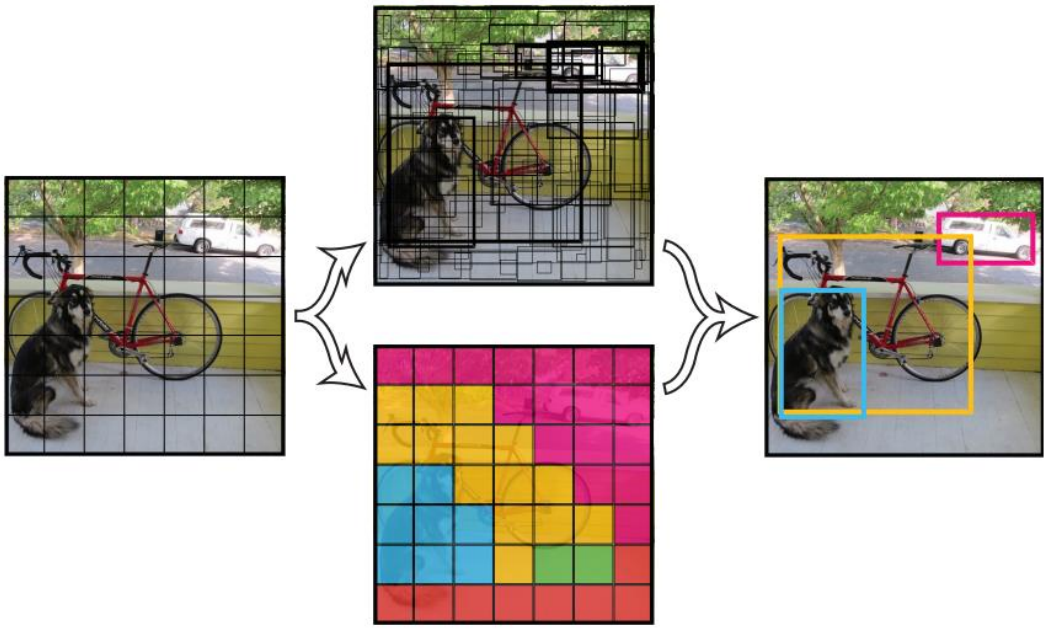


Figure V.4 : Principe de fonctionnement de l'architecture YOLO [83].

La figure V.5 représente l'architecture détaillée du système YOLO utilisée pour la reconnaissance d'objets.

Name	Filters	Output Dimension
Conv 1	7 x 7 x 64, stride=2	224 x 224 x 64
Max Pool 1	2 x 2, stride=2	112 x 112 x 64
Conv 2	3 x 3 x 192	112 x 112 x 192
Max Pool 2	2 x 2, stride=2	56 x 56 x 192
Conv 3	1 x 1 x 128	56 x 56 x 128
Conv 4	3 x 3 x 256	56 x 56 x 256
Conv 5	1 x 1 x 256	56 x 56 x 256
Conv 6	1 x 1 x 512	56 x 56 x 512
Max Pool 3	2 x 2, stride=2	28 x 28 x 512
Conv 7	1 x 1 x 256	28 x 28 x 256
Conv 8	3 x 3 x 512	28 x 28 x 512
Conv 9	1 x 1 x 256	28 x 28 x 256
Conv 10	3 x 3 x 512	28 x 28 x 512
Conv 11	1 x 1 x 256	28 x 28 x 256
Conv 12	3 x 3 x 512	28 x 28 x 512
Conv 13	1 x 1 x 256	28 x 28 x 256
Conv 14	3 x 3 x 512	28 x 28 x 512
Conv 15	1 x 1 x 512	28 x 28 x 512
Conv 16	3 x 3 x 1024	28 x 28 x 1024
Max Pool 4	2 x 2, stride=2	14 x 14 x 1024
Conv 17	1 x 1 x 512	14 x 14 x 512
Conv 18	3 x 3 x 1024	14 x 14 x 1024
Conv 19	1 x 1 x 512	14 x 14 x 512
Conv 20	3 x 3 x 1024	14 x 14 x 1024
Conv 21	3 x 3 x 1024	14 x 14 x 1024
Conv 22	3 x 3 x 1024, stride=2	7 x 7 x 1024
Conv 23	3 x 3 x 1024	7 x 7 x 1024
Conv 24	3 x 3 x 1024	7 x 7 x 1024
FC 1	-	4096
FC 2	-	7 x 7 x 30 (1470)

Figure V.5 : Architecture détaillée du YOLO V1.

Plusieurs versions du YOLO ont été proposées à ce jour, on trouve : YOLO V1, YOLO V2, VOLO V3, Tiny YOLO V3. Chaque version a pour but d'améliorer les performances du modèle en matière de rapidité et taux de classification. Dans ce document, nous allons travailler sur la version « 1 » du modèle YOLO.

Tous les modèles de base de YOLO ont été entraînées sur la base de données de COCO, mais il existe dans la littérature d'autres modèles entraînés sur d'autres bases de données telles que ImageNet de Google.

V.4. Description de la méthode proposée

La méthode proposée est basée sur l'utilisation de l'architecture YOLO V1 présentée précédemment, nous proposons l'utilisation de l'affinement (*fine-tuning*) du modèle YOLO V1 et de l'adapter au problème de reconnaissance d'activités humaines.

À l'origine, ce modèle est capable de faire la classification de 80 objets différents, entraînés sur la base de données de COCO. Le principe de notre méthode est d'utiliser l'architecture du modèle YOLO pour refaire l'apprentissage sur les bases de données de reconnaissance d'activités humaines.

Notre approche permet une reconnaissance d'activités image par image, c'est-à-dire que la reconnaissance va se faire en temps réel et chaque image du flux vidéo est classée séparément.

L'organigramme de la méthode proposée est représenté sur la figure V.6 :

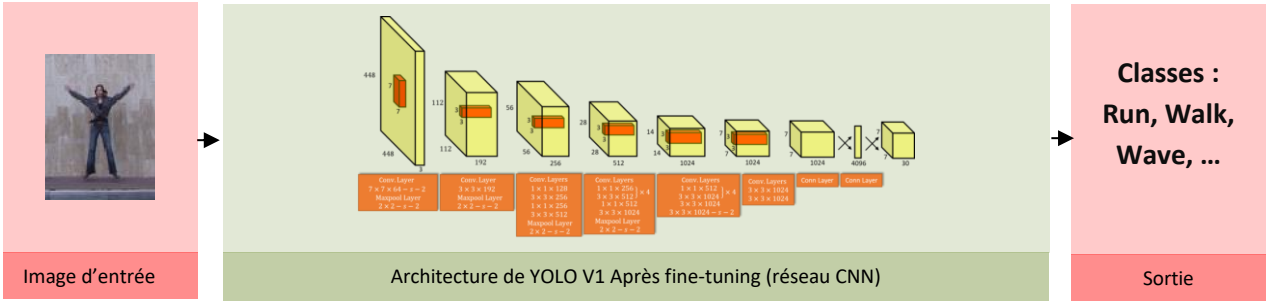


Figure V.6 : Organigramme de la méthode proposée.

Comme présenté précédemment, les sorties du modèle YOLO sont les coordonnées des zones englobantes (*bounding boxes*) et les probabilités de classification de chaque objet correspondant, pour transformer cette architecture pour la reconnaissance d'activités humaines, nous avons supprimé les deux dernières couches entièrement connectées du modèle YOLO, ensuite nous avons inséré une couche entièrement connectée et une couche *Softmax* suivie d'une couche de classification, donc la sortie de notre nouvelle architecture est les classes d'appartenance des activités dans chaque image du flux vidéo.

La dernière étape est le fine-tuning du modèle par réentrainement en utilisant la nouvelle base de données sur la reconnaissance d'activités humaines.

Le but de la méthode proposée est la reconnaissance des activités image par image en temps réel. Pour adapter notre méthode pour la reconnaissance des activités dans les

séquences vidéo pour les applications d'indexation et de recherches automatisée par exemple, nous proposons un protocole de fusion des résultats de reconnaissance de chaque image.

Le protocole de fusion proposé est appliqué lors de la phase de décision, c'est-à-dire la fusion des scores est réalisée en dehors du réseau CNN et après que ce dernier délivre ses résultats pour toutes les images de la séquence vidéo.

Le protocole proposé est composé de deux étapes (figure V.7), la première étape est un filtre de fiabilité, si le score de classification est inférieur à un seuil T, la classification est rejetée. Et une deuxième étape de recherche de la classe pertinente dans la séquence par l'utilisation d'histogramme :la classe qui correspond à la valeur maximale de l'histogramme des classes de toutes les images après seuillage, représente la classe d'appartenance de la séquence vidéo.

Le diagramme suivant illustre le protocole de fusion proposé :

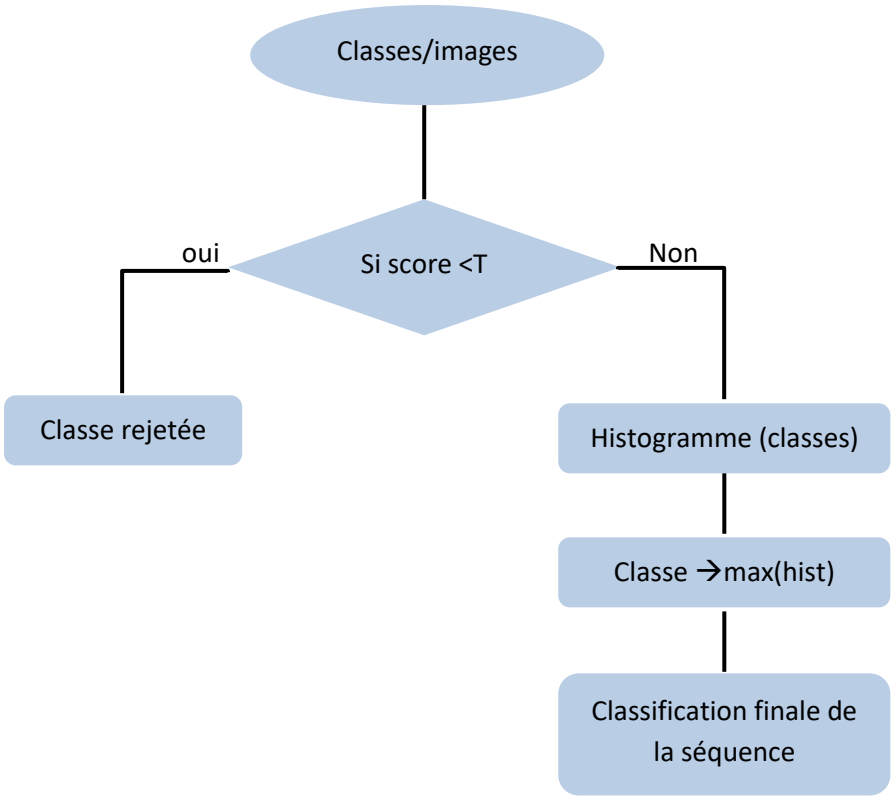


Figure V.7 : Organigramme du Protocole de fusion proposé.

V.5. Résultats expérimentaux

V.5.1. Protocole de test

Pour évaluer les performances de la technique proposée, nous avons utilisé la base de données de KTH qui reste toujours un challenge dans le domaine de reconnaissance d'activités humaines.

Pour faire une reconnaissance image par image en temps réel, on a ajouté une nouvelle classe à la base de données KTH nommée « **no action** » qui représente les images qui ne contiennent aucun sujet. Ainsi, la base de données finale va contenir les classes suivantes : No Action, Run, Jogging, Walk, Boxing, waving et hand-clapping.

Nous avons divisé la base de données KTH en deux sous-ensembles : 70% pour l'entrainement et 30% pour le test. Dans chaque classe, Nous avons 70 séquences vidéo pour l'entrainement et 30 séquences vidéo pour le test.

Le tableau V.1 montre la subdivision de la base de données KTH utilisée dans la simulation :

	Séquences	**Images**
Entrainement (70%)	420	33600
Test (30%)	180	14320

Tableau V.1 : Sous-ensembles utilisés dans l'apprentissage.

V.5.2. Discussion des résultats

a. Reconnaissance des activités image par image

Nous avons réalisé plusieurs simulations en utilisant en premier temps l'apprentissage par transfert (*transfert learning*), ensuite nous avons réalisé un affinement complet (*fine-tuning total*) du modèle de base par la mise à jour des poids de toutes les couches dans l'architecture YOLO. Le tableau V.2 montre les résultats obtenus :

Niveau du Fine-tuning	**Taux de reconnaissance**
Transfert learning	52.00%
Fine-tuning Total	**82.00%**

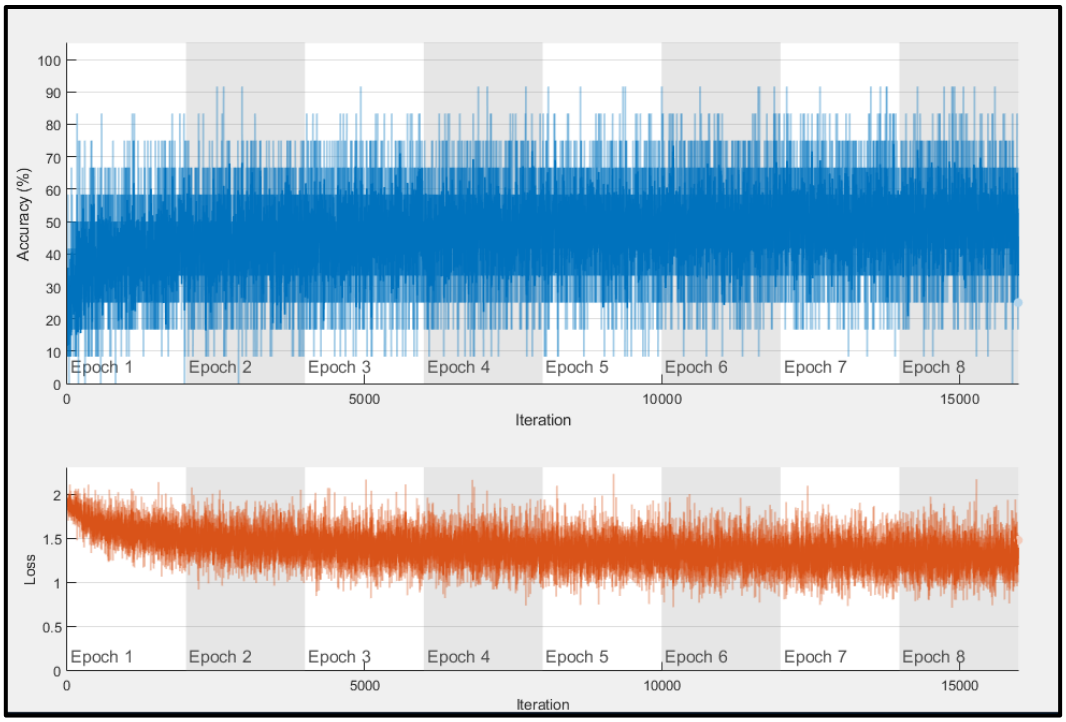
Tableau V.2 : Résultats de reconnaissance en utilisant différents niveaux de fine-tuning.

Les taux de reconnaissance du tableau V.2 montrent que les meilleures performances ont été obtenues en utilisant un fine-tuning total des poids du modèle YOLO de base, cela est expliqué par la différence entre le problème de base pour lequel le modèle a été entrainé, et le problème de reconnaissance d'activités humaines. À l'origine, le but du modèle été de fournir les cordonnées des zones englobantes ainsi que les scores des classes dans ces zones, par contre dans notre problème, le modèle est entrainé pour fournir directement les classes d'appartenance des activités dans les images. En d'autres termes, les cartes de descripteurs générées par le modèle de base ne sont pas adaptées pour le problème de reconnaissance d'activités humaines, cependant, l'architecture du modèle a permis des résultats très encourageants avec un taux de reconnaissance de **82%** des images.

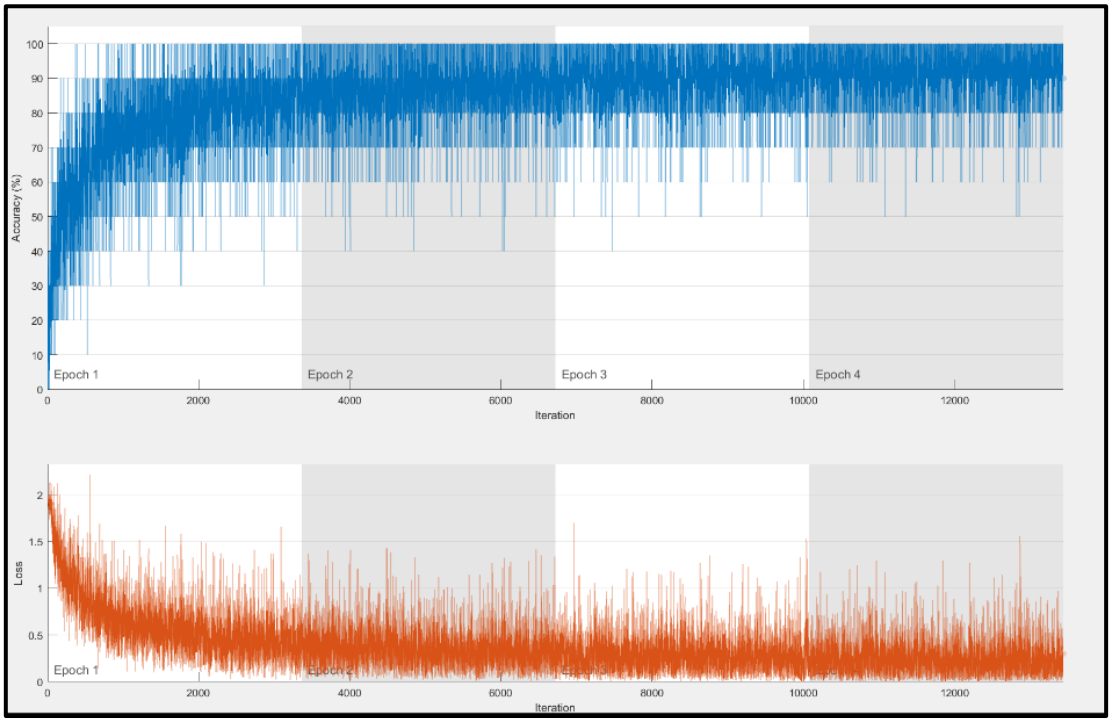
Des exemples de reconnaissance tirés directement des vidéos sont représentées dans la figure V.8 suivante :

Figure V.8 : Exemple de reconnaissance en utilisant la base de données de KTH.

Les courbes d'apprentissage en utilisant le transfert learning et le fine-tuning total sont représentées sur la figure V.9 :



a



b

Figure V.9 : Courbes d'apprentissages en utilisant la base de données KTH.

a) Transfert learning, b) fine-tuning total

Les courbes d'apprentissage sur la figure V.9 montrent clairement que le fine-tuning total a permis d'obtenir un modèle plus performant avec un taux t très élevé et une erreur très petite. Par contre, le transfert learning montre un comportement instable avec des taux très faible et une erreur relativement élevée.

Les résultats de reconnaissance obtenus en utilisant le fine-tuning total sont détaillés dans la matrice de confusion de la figure V.10.

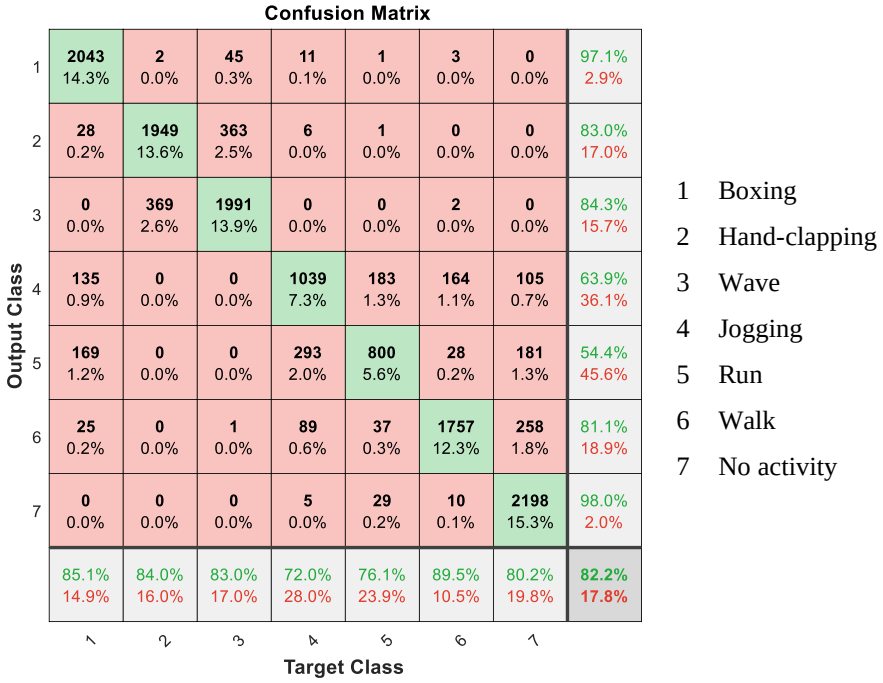


Figure V.10 : Matrice de confusion en utilisant le fine-tuning total sur la base de données KTH.

La matrice de confusion lors de l'utilisation du fine-tuning Total, montre que le modèle généré a permis d'obtenir des taux de reconnaissance supérieurs à 72% pour toutes les activités, cela confirme la stabilité et la robustesse du modèle.

La courbe Roc e la figure V.11, confirme la fiabilité du modèle final réalisé par le fine-tuning total du modèle de base YOLO V1. Les courbes ROC de toutes les activités sont proches de 1, et la valeur moyenne de la surface sous la courbe (Area Under the Curve) des activités AUC est de **0.9812**

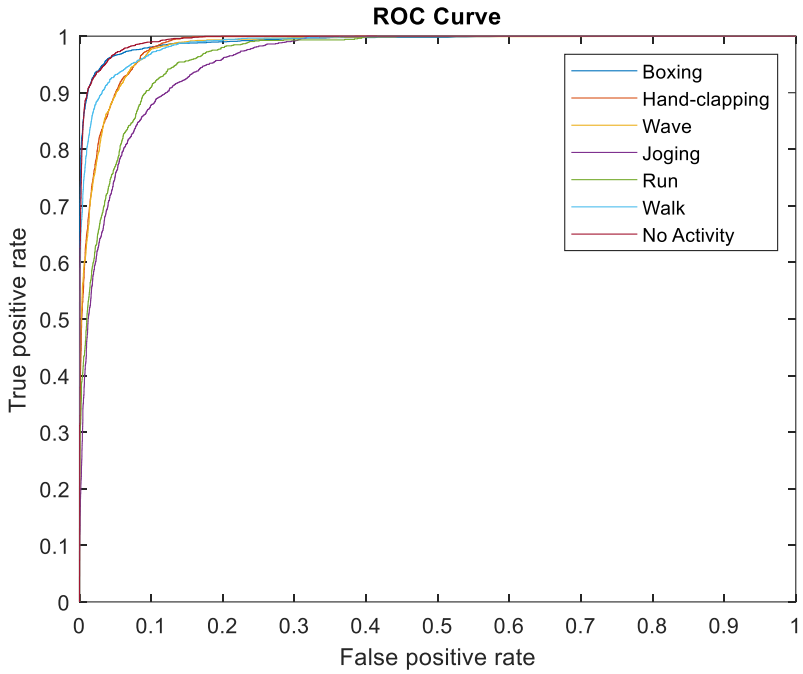


Figure V.11 : La courbe ROC du modèle YOLO fine-tuné en utilisant la base de données KTH.

b. Reconnaissance des activités dans les séquences vidéo

Pour adapter notre méthode pour la reconnaissance des activités dans les séquences vidéo, nous avons proposé un protocole de fusion des classes obtenues par toutes les images de la séquence vidéo.

Nous avons réalisé plusieurs tests en utilisant différentes valeurs de seuil T des scores, le tableau V.3 montre les résultats obtenus :

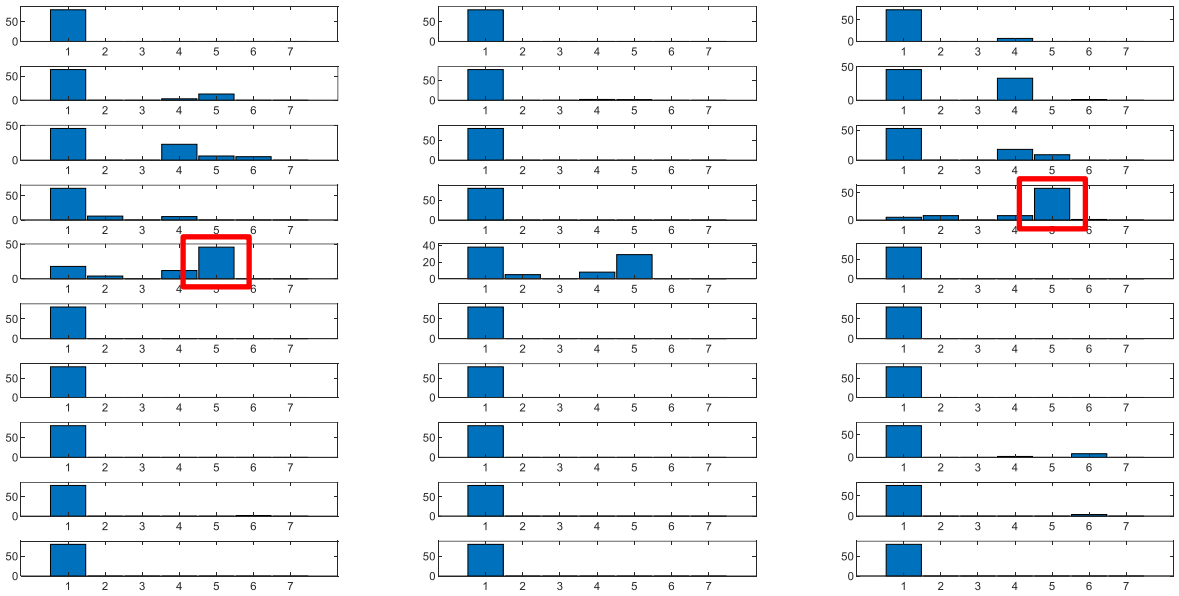
Seuil T	Taux de reconnaissance
0.35	93.88%
0.40	93.88%
0.45	**94.44%**
0.50	**94.44%**
0.55	93.88%
0.6	93.88%

Tableau V.3 : Taux de reconnaissance en utilisant différentes valeurs de seuil T.

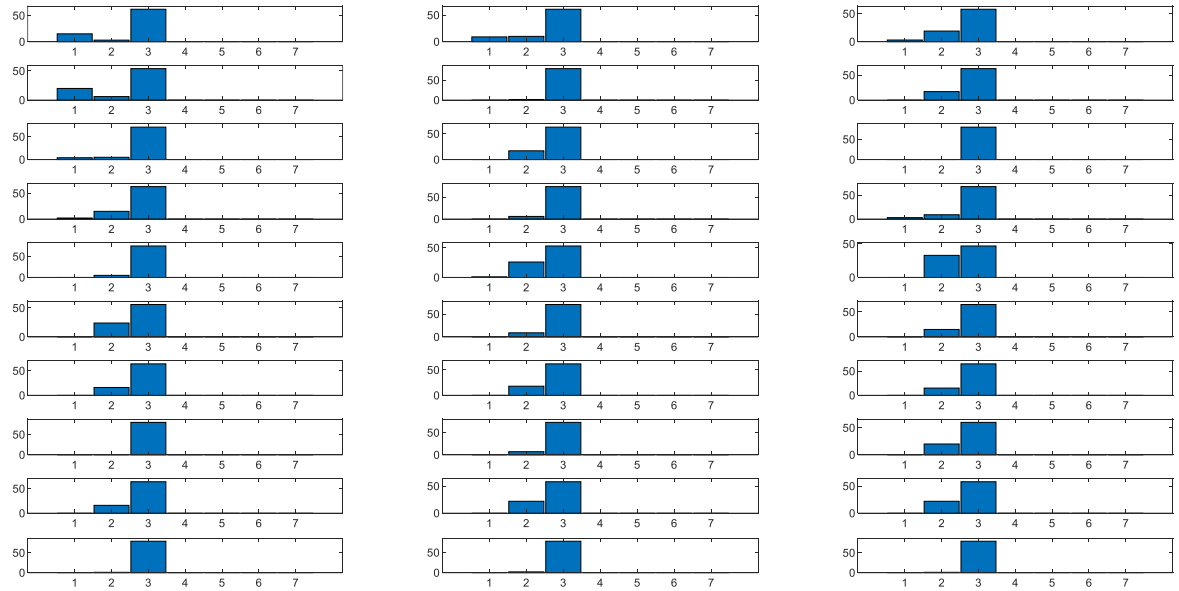
Les résultats du tableau V.3, montrent que les meilleurs taux de reconnaissance ont été obtenus par l'utilisation d'un seuil **0.45<T<0.5**, ce résultat signifie que les images avec un

score inférieur à 0.5 (un faible score signifie une faible fiabilité de reconnaissance) sont rejetées et ne sont pas considérées dans le calcul des histogrammes des classes. Les résultats obtenus confirment que le seuil T a une grande importance dans l'opération de fusion, il permet d'améliorer la fiabilité de la reconnaissance dans la séquence vidéo.

La figure V.12 montre les histogrammes du protocole de fusion utilisé sur toutes les séquences vidéo de test pour les deux activités Boxing et Waving :



1 Boxing, 2 Hand-clapping, 3 Wave, 4 Jogging, 5 Run, 6 Walk, 7 No activity (a)



1 Boxing, 2 Hand-clapping, 3 Wave, 4 Jogging, 5 Run, 6 Walk, 7 No activity (b)

Figure V.12 : Histogrammes des classes des séquences vidéo de test pour les deux activités : a) Boxing, b) Waving.

La figure V.12, montre un exemple des histogrammes de classes dans les séquences vidéo de test, obtenues en utilisant notre protocole de fusion. Dans cette figure, nous avons les histogrammes de 30 séquences vidéo de chaque activité (Boxing, Waving). La classe qui correspond à la valeur maximale de chaque histogramme, représente la classe d'appartenance de la séquence vidéo. La figure (a) montre toutes les histogrammes de l'activité Boxing, avec en rouge deux séquences mal classées, par contre pour l'activité Waving en (b) la classification est parfaite.

Le tableau V.4 montre les performances de notre approche par rapport aux autres techniques présentées dans la littérature.

Méthodes	Taux de reconnaissance
Wong and Cipolla. 2007 [88]	86.62%
Niebles et al. 2007 [89]	83.33%
Laptev et al. 2008 [46]	92.10%
Schuldt et al. 2004 [45]	71.70%
Dollar et al. 2005 [27]	81.20%
Bo Chen et al. 2010 [90]	91.13%
Vivek et al. 2015 [91]	93.96%
Lin Sun et al. 2014 [92]	93.10%
Moez B et al. 2015 [15]	94.39%
Méthode proposée (BSTM+CNN)	**92.50%**
Méthode proposée (YOLO+ Fine-tuning total)	**94.44%**

Tableau V.4 : Taux de reconnaissance en utilisant la base de données de KTH.

Les résultats du tableau V.4 montrent que la méthode proposée donne de meilleurs résultats comparés aux techniques de l'état de l'art et aussi aux approches que nous avons proposées dans les chapitres précédents.

Les détails des scores de reconnaissance sont donnés par la matrice de confusion de la figure V.13. On peut remarquer que la majorité des fausses classifications sont liées aux activités « run » et « jogging », cela peut être expliqué par la ressemblance des deux activités. Malgré cette confusion, les taux de reconnaissance des deux activités restent très élevés, de l'ordre de **83.3%** et **96.7%** respectivement.

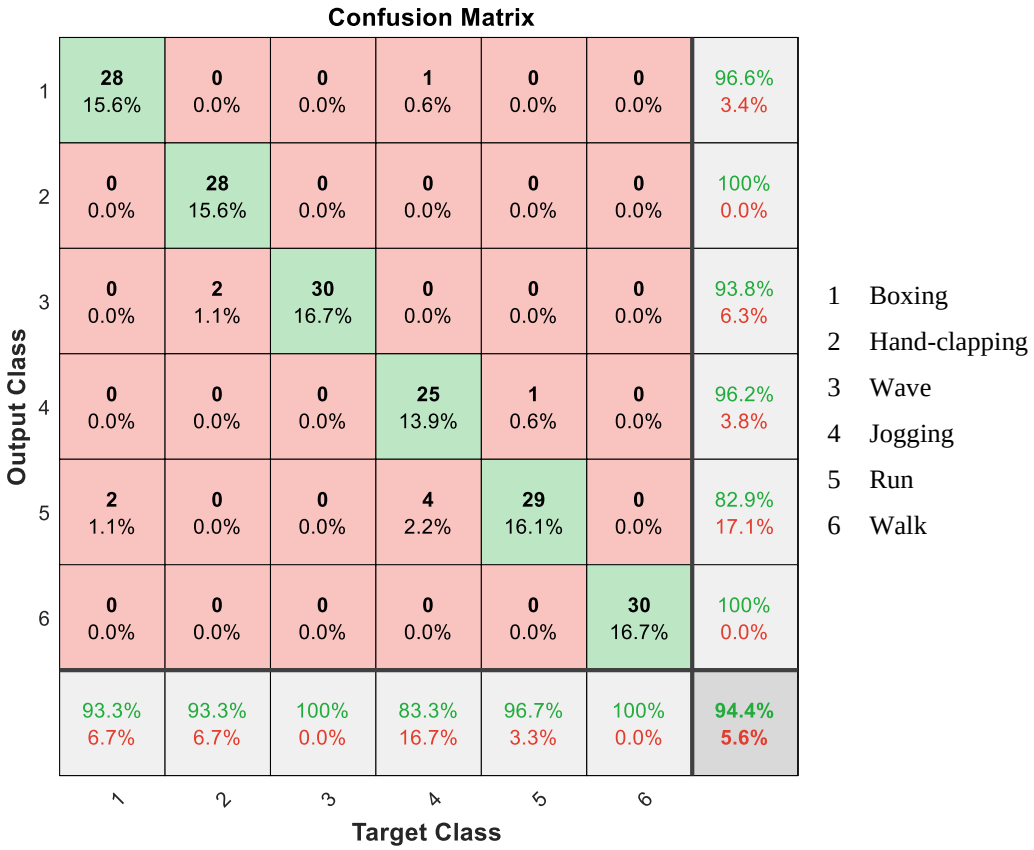


Figure V.13 : Matrice de confusion en utilisant la base de données KTH.

c. Application sur la base de données de Weizmann

Les bons résultats obtenus par notre approche utilisant le modèle YOLO, ainsi que sa simplicité et sa rapidité, nous ont confortés pour étendre son application à un environnement réel. Nous avons donc réalisé une interface graphique qui permet la classification des activités dans les séquences vidéo et aussi la reconnaissance en temps réel en utilisant la caméra. Cette interface offre la possibilité d'observer le comportement du modèle d'apprentissage profond sur chaque image en affichant le score et la classe, ainsi à la fin l'interface calcule la classe d'appartenance de la séquence vidéo.

Un aperçu de l'interface réalisée est présenté sur la figure V.14 :

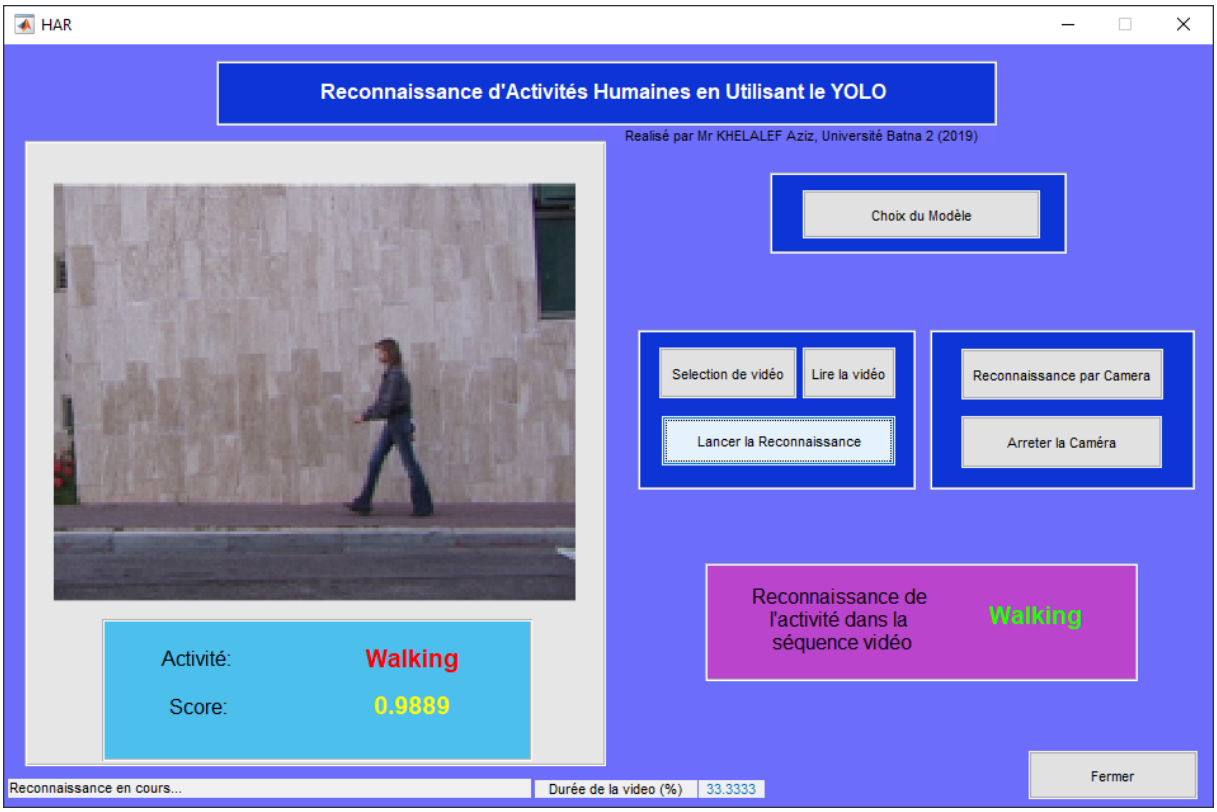


Figure V.14 : Interface graphique de reconnaissance.

Pour tester la robustesse de la méthode proposée, nous avons utilisé l'interface de simulation et la base de données de Weizmann qui n'a pas été considérée durant la phase d'apprentissage du modèle final.

Lors de la simulation, nous avons considéré que les activités communes entre les deux bases de données : Run, Walk, Wave1 et Wave2. Dans un premier temps, nous avons utilisé toutes les séquences vidéo de chaque activité dans le test (10 activités Run, 10 activités Walk, 9 activités Wave1 et 9 activités Wave2). Par la suite, nous avons utilisé l'ensemble de test des deux chapitres précédents.

Des exemples de reconnaissance tirés de l'interface de simulation sont représentés sur la figure V.15:

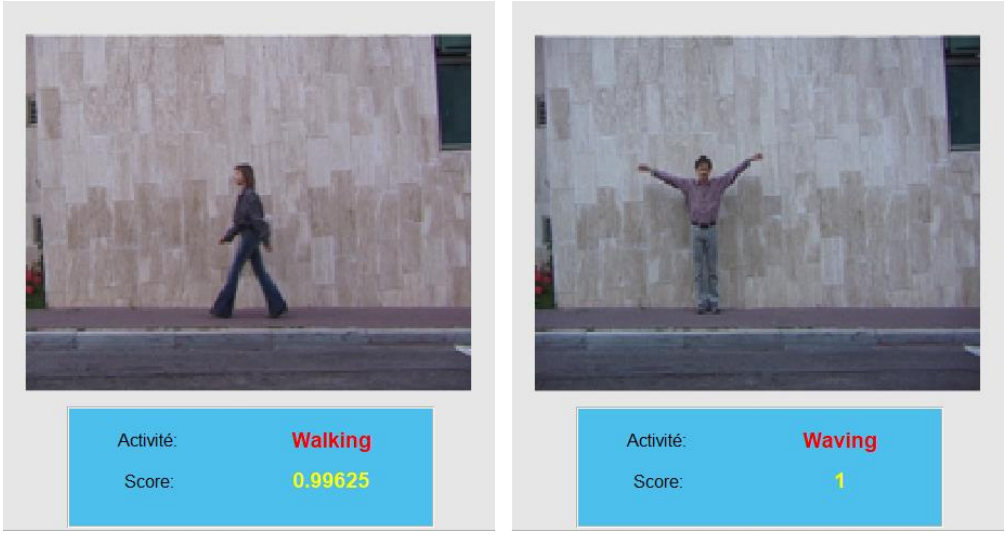


Figure V.15 : Exemple de simulation en utilisant la base de données de Weizmann et le modèle YOLO fine-tuné.

Les résultats expérimentaux dans le tableau V.5, montrent que le modèle YOLO fine-tuné en utilisant la base de données KTH a permis un taux de reconnaissance de **89.5% (34/38 activités)** sur la base de données de Weizmann en utilisant toutes les séquences des quatre (4) activités et un taux de reconnaissance de **93.8% (15/16 activités)** lors de l'utilisation de l'ensemble de test seulement.

Méthodes	Taux de reconnaissance
Boiman and Irani 2006 [**85**]	97.5% (9 activités)
Scovanner et al. 2007 [**29**]	82.6% (10 activités)
Wang and Suter 2007 [**86**]	97.8% (10 activités)
Kellokumpu et al 2008 [**87**]	97.8% (10 activités)
Kellokumpu et al. 2009 [**37**]	98.7% (9 activités)
Hafiz Imtiaz et al. 2015 [**44**]	100% (10 activités)
Tasweer et al. **2015 [77**]	92.25% (10 activités)
Tushar et al. 2015 [14]	100% (5 activités)
Méthode proposée (DCT+SVM)	**92.50% (10 activités)**
Méthode proposée (BSTM+CNN)	**98% (10 activités)**
Méthode propose (YOLO+fine-tuning total)	**89.5% (4 activités + toutes les séquences)**
Méthode propose (YOLO+fine-tuning total)	**93.8% (4 activités + ensemble test)**

Tableau V.5 : Taux de reconnaissances en utilisant la base de données de Weizmann

La matrice de confusion en utilisant la totalité de la base de données de Weizmann (figure V.16) montre que le taux de reconnaissance pour l'activité Run est de **70% (7/10 séquences)**, les **30%** d'erreur sont classés comme activité Walk qui sont des activités très similaires. Ainsi, le modèle a permis un taux de classification de **100%** pour l'activité Walk et un taux de **94.4% (17/18 séquences)** pour l'activité Wave.

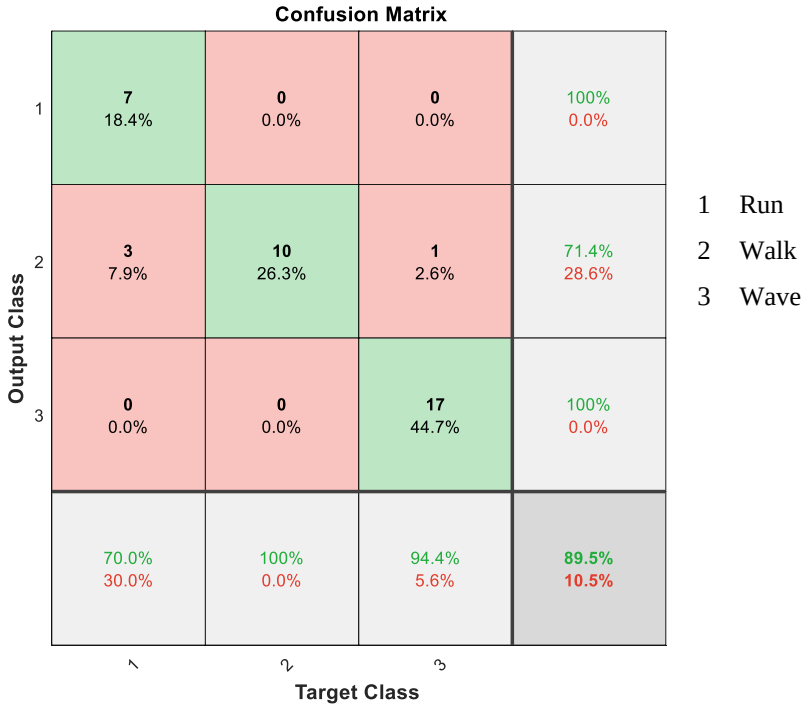


Figure V.16 : Matrice de confusion en utilisant la base de données de Weizmann et le modèle YOLO fine-tuné.

Lors de l'utilisation de l'ensemble de test seulement dans la reconnaissance, la méthode proposée basée sur le modèle YOLO fine-tuné a permis un bon taux de classification de 93.8% (15/16 séquences), les résultats détaillés de la classification sont donnés par la matrice de confusion de la figure V.17 :

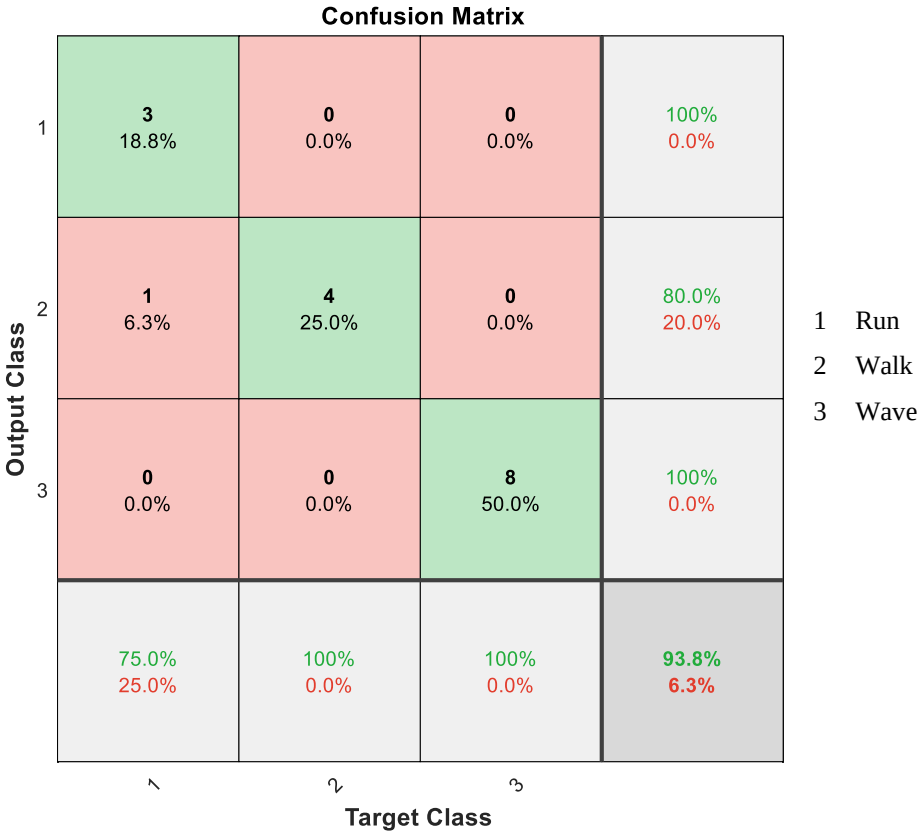


Figure V.17 : Matrice de confusion lors de l'utilisation de l'ensemble test de la base de données de Weizmann.

d. Application sur les vidéos YouTube

Après les résultats encourageants lors de l'utilisation de notre modèle fin-tuné sur la base de données de Weizmann, nous avons réalisé plusieurs tests avec le simulateur en utilisant des vidéos réelles tirées directement du YouTube, nous avons choisi des vidéos qui contiennent les mêmes activités qui ont servi à l'entrainement du système

La figure V.18 montre des exemples de reconnaissances en utilisant notre interface de simulation :

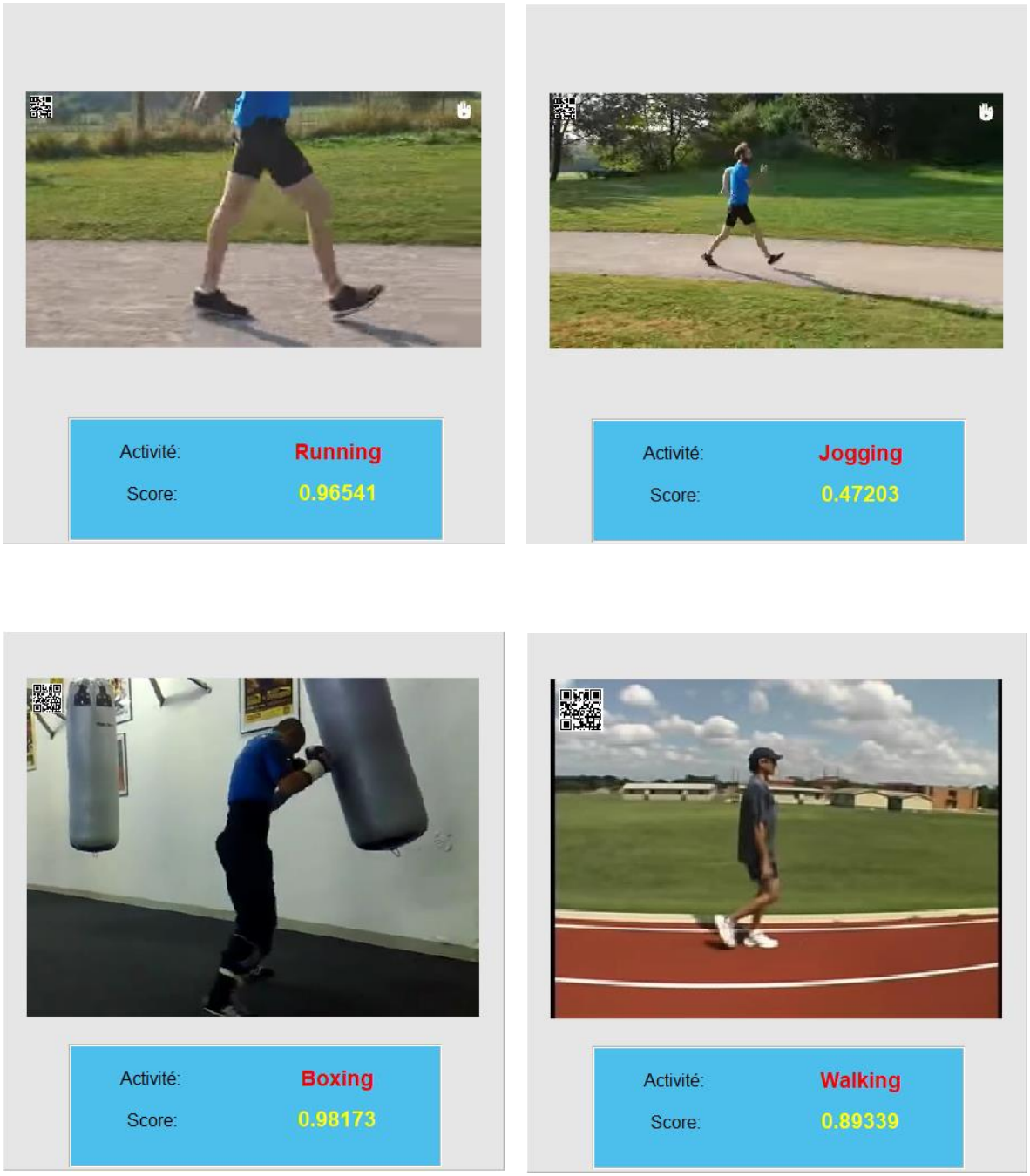


Figure V.18 : Résultats de simulation en utilisant des vidéos YouTube.

Les résultats de reconnaissance sur les vidéos réelles tirées de YouTube ont montré que le modèle YOLO fine-tuné a permis une bonne reconnaissance des activités, par exemple la figure V.18 (a) montre que la méthode proposée a donné de bonnes performances même si le corps humain n'est pas complétement visible sur les images.

Notre approche a permis de reconnaitre les diverses activées utilisées dans l'apprentissage. Ces résultats nous laissent penser qu'avec l'utilisation d'autre base de données plus larges telle que UCF-101 et des vidéos tirées de YouTube par exemple, la

méthode proposée pourrait atteindre des taux de reconnaissance supérieurs même dans les vidéos avec des activités complexes.

e. Application en temps réel en utilisant la Camera

Afin d'appliquer notre méthode dans des conditions de reconnaissance d'activités en temps réel, nous avons équipé notre simulateur d'une Webcam. Ainsi, la reconnaissance est réalisée en temps réel image par image et les résultats de reconnaissance sont affichées directement sur l'interface à chaque image.

La figure V.19 suivante montre des exemples de reconnaissance en utilisant la caméra, tirées directement de l'interface de simulation :

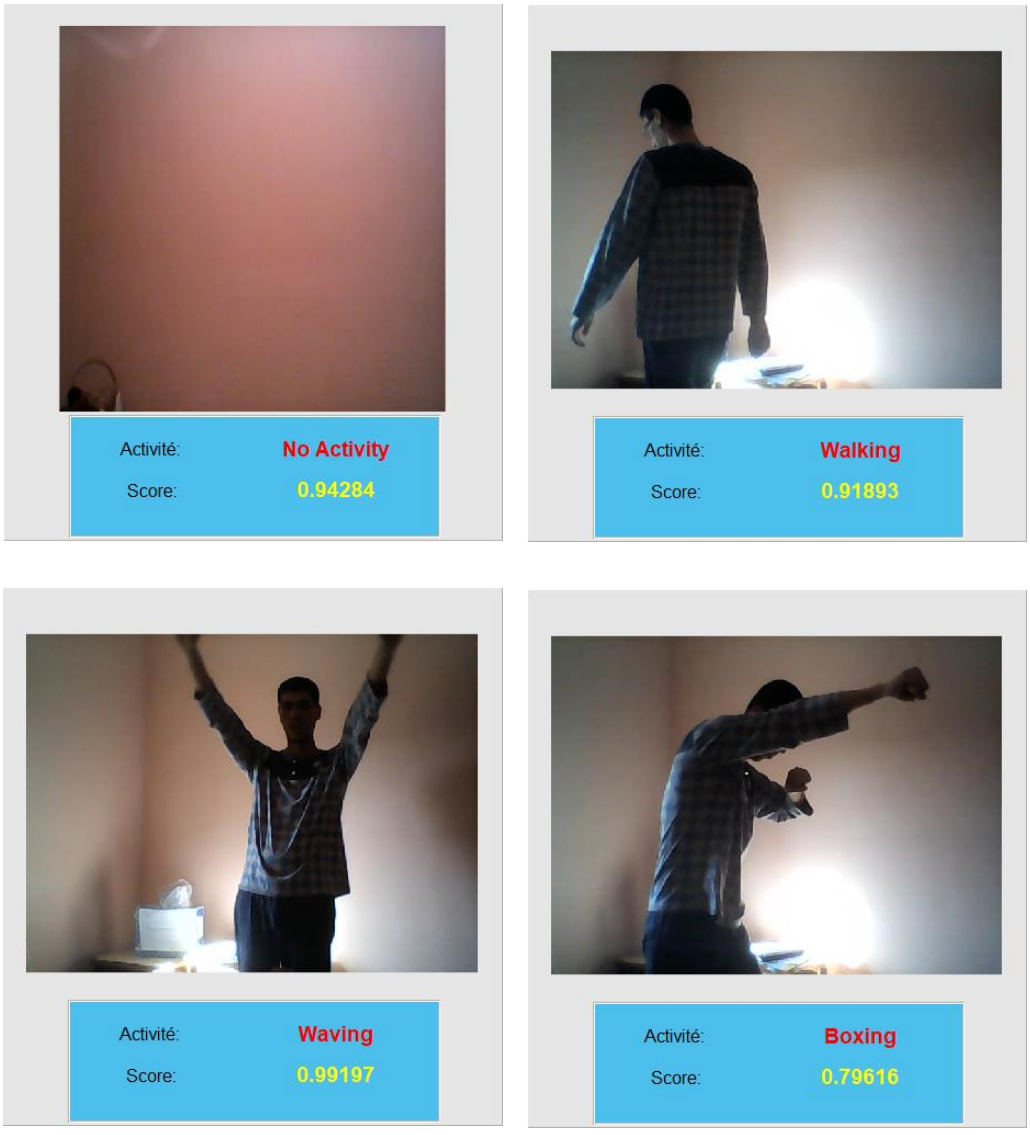


Figure V.19 : Exemple de simulation en temps réel en utilisant la caméra.

V.6. Conclusion

Dans ce chapitre, nous avons proposé une nouvelle méthode de reconnaissance d'activités humaines en utilisant l'architecture YOLO. La méthode est basée sur le fine-tuning des poids de toutes les couches de l'architecture du modèle pour la reconnaissance image par image en temps réel, et un protocole de fusion proposé pour la reconnaissance des activités dans les séquences vidéo.

Le choix du modèle YOLO est lié aux performances exceptionnelles atteintes par ce dernier : les meilleurs taux de reconnaissance sur les bases de données de référence dans le domaine de reconnaissance d'objets, et la rapidité du modèle qui est basé sur l'utilisation d'un seul passage sur l'image d'entrée.

Nous avons proposé un protocole de fusion au niveau de décision, c'est-à-dire après la classification de toutes les images de la séquence vidéo par le modèle YOLO fine-tuné. Le protocole est composé d'une étape de fiabilité basée sur le rejet des classes avec un score inférieur à un seuil T, et une étape de décision basée sur l'utilisation des histogrammes des classes de toutes les images de la séquence vidéo.

Les résultats expérimentaux ont montré que la méthode proposée a donné des résultats de reconnaissance très performants lors de la reconnaissance image par image avec un taux de reconnaissance de **82.2%** en utilisant la base de données de KTH.

Notre approche a donné aussi un taux de reconnaissance de **94.44%** surpassant ainsi les techniques de la littérature et aussi les méthodes basées sur la DCT et les BSTM que nous avons proposées dans un premier temps.

Ces performances ont été confirmées par l'utilisation de la base de données de Weizmann où nous avons obtenu un taux de reconnaissance de **93.8%** lors de l'utilisation de l'ensemble de test et **89.5%** lors de l'utilisation de la base de données globale des 4 activités Run, Walk, Wave1 et Wave 2.

Pour tester l'efficacité de notre méthode dans les conditions réelles, nous avons conçu une interface graphique pour la simulation de la reconnaissance des activités, cette dernière a permis la reconnaissance image par image en temps réel en utilisant la caméra, et aussi la reconnaissance des activités dans les séquences vidéo.

Nous avons utilisé l'interface de simulation pour tester notre technique sur des vidéos réelles tirées de YouTube et aussi en utilisant la caméra, les résultats de reconnaissance ont montré que notre méthode arrive à identifier des activités utilisées dans l'apprentissage sur les séquences réelles plus complexes que celles de KTH ou Weizmann (fond variable, déplacement du point de vue et changement de luminosité).

La méthode proposée reste performante pour l'ensemble de tests réalisés, elle donne des taux de reconnaissance supérieurs pour toutes les bases de données utilisées et aussi pour des séquences vidéo réelles.

L'utilisation de l'architecture YOLO, nous a permis de profiter de la rapidité du modèle pour proposer une technique de reconnaissance image par image et en temps réel.

Enfin, nous sommes convaincus que la méthode proposée peut donner des résultats de reconnaissance encore meilleurs si elle est appliquée à des base de données plus larges et plus complexes tel que UCF-101. Cela nécessitera certes des machines plus puissantes avec des cartes graphiques de dernière génération et des disques de stockage de grandes capacités et aussi du temps pour aboutir optimale configuration du modèle.

Conclusion Générale

« Ce n'est pas dans la science qu'est le bonheur, mais dans l'acquisition de la science. »

Edgar Allan Poe
Artiste, écrivain, Poète, Romancier (1809 - 1849)

Conclusion Générale.

La reconnaissance d'activités humaines est un axe de recherche d'actualité dans le domaine de la vision par ordinateur. Donner à la machine le pouvoir d'analyser et d'interpréter les activités réalisées par une personne dans une scène vidéo permet l'ouverture de plusieurs champs d'applications. Donner ainsi à l'humain le pouvoir de contrôler l'énorme volume massif des données vidéo enregistrées et diffusés chaque jour dans le monde.

Dans cette thèse, nous avons essayé de présenter le domaine de la reconnaissance d'activités humaines en utilisant les descripteurs spatio-temporels 2D/3D. Nous avons présenté les problématiques et les contraintes rencontrées dans cet axe de recherche, ainsi que les concepts conventionnels et récents proposés dans la littérature pour les surmonter.

Nous avons consacré la première partie de notre thèse à la présentation de l'état de l'art du domaine de reconnaissance d'activités humaines. Nous avons essayé d'exposer les techniques conventionnelles catégorisées selon le type de descripteurs utilisés, en commençant par présenter les techniques basées sur les descripteurs locaux en utilisant les concepts développés dans le domaine de la reconnaissance d'objets. Nous avons ensuite exposé les techniques basées sur les descripteurs globaux basés sur l'extraction de la forme et de la dynamique globale du corps humain, par la suite. Nous avons montré par la suite les techniques basées sur la modélisation du corps humain et son évolution dans l'espace.

Nous avons présenté également dans cette partie les différentes méthodes de classification utilisées par les auteurs du domaine. Une étude détaillée sur les réseaux de neurones artificiels et l'apprentissage profond (*deep learning*) été également présentée. À la fin de cette partie, nous avons présenté les différentes techniques de reconnaissance d'activités humaines basées sur l'apprentissage profond.

Orienté par les travaux présentés dans la littérature, ainsi que par les progrès considérables dans le domaine de machine learning, et dans un but de contribuer à résoudre les problématiques de ce champ de recherche, dans la deuxième partie de cette thèse, nous avons présenté en détail notre contribution dans le domaine en proposant quatre techniques de reconnaissance d'activités humaines.

Nous avons proposé une nouvelle technique de reconnaissance d'activités humaines basée sur l'extraction des skeletons et la transformée en cosinus discrète DCT pour le calcul des caractéristiques, et la SVM pour la classification des activités. L'étude comparative a montré que la méthode proposée a donné un taux de reconnaissance de **92.5%** lors de l'utilisation de dix (10) activités dans la base de données de Weizmann et un taux de **100%** lors de l'utilisation de seulement neuf (9) activités. La comparaison avec les résultats de la littérature montre que notre méthode a donné des résultats comparables et satisfaisants.

Le fait que la méthode proposée est basée sur le calcul des cartes spatio-temporelles, cela a pu rendre son utilisation impossible pour la reconnaissance des activités image par image en temps réel. Pour remédier à ce problème, nous avons présenté une alternative à cette technique en utilisant les silhouettes et la transformée en cosinus discrète DCT pour l'extraction des caractéristiques et les réseaux de neurones artificiels multicouches RBF pour la reconnaissance. La méthode proposée a donné un taux de reconnaissance très intéressant de **99%** sur la base de données de Weizmann.

Le point faible de ces deux premières techniques réside dans l'algorithme de classification utilisé. En effet, d'un côté, le SVM souffre de la baisse de performance lorsque le nombre de classes ou la taille des bases de données augmente. De l'autre côté, le réseau RBF a tendance à tomber dans les minimas locaux et le sur-apprentissage, en plus de la limitation du nombre de neurones dans la couche cachée. Pour cela l'utilisation de nouvelles bases de données plus complexes telle que KTH s'est avérée impossible.

Pour pallier ces problèmes, nous avons proposé une nouvelle méthode de reconnaissance d'activités humaines basée sur un nouveau descripteur appelé BSTM (*Binary Space-time Maps*) qui est la combinaison de l'information spatiale et l'information temporelle dans un intervalle de temps, et les réseaux de neurones à convolution CNN pour la reconnaissance des activités.

Les résultats expérimentaux ont montré que la nouvelle méthode proposée a donné des résultats très compétitifs en surpassant les techniques présentées dans la littérature et en délivrant des performances équivalentes aux techniques récentes basées sur l'apprentissage profond avec des taux de reconnaissance de **98%** sur la base de données de Weizmann, **100%** sur la base de données Keck Gesture Database et **92.5%** sur la base de données KTH.

Malgré les résultats atteints, notre méthode s'est révélée incapable de faire la reconnaissance des activités image par image en temps réel à cause des BSTM qui sont calculés par l'empilement de plusieurs images dans un intervalle de temps. Pour pallier ce problème, nous avons proposé une quatrième méthode de reconnaissance d'activités humaines totalement automatisée en utilisant l'apprentissage profond.

A cet égard, nous avons proposé une méthode basée sur l'apprentissage par transfert en utilisant le modèle YOLO (*you only look once*) entrainé initialement pour la reconnaissance d'objet, nous avons utilisé un fine-tuning totale du modèle et un nouvel apprentissage sur la base de données KTH. Les résultats expérimentaux ont montré que la méthode proposée a donné des résultats de reconnaissance image par image très performante avec un taux de reconnaissance de **82%** sur la base de données KTH.

Encouragé par les résultats obtenus lors de la reconnaissance image par image, nous avons adapté cette technique pour la reconnaissance des activités dans les séquences vidéo, pour cela, nous avons proposé un nouveau protocole de fusion des résultats de reconnaissance de chaque image pour délivrer à la fin l'activité contenue dans la séquence vidéo. Les résultats expérimentaux ont montré que la technique a donné des très bons taux de reconnaissance de **94.44%** en surpassant toutes les techniques présentées dans la littérature. Nous pensons que notre méthode est simple et rapide, parce que n'elle ne nécessite aucun prétraitement des images, l'entrée du modèle final reposant sur les images brutes du flux vidéo.

Dans le futur, et comme perspective aux travaux que nous avons déjà commencés, nous prévoyons la réalisation d'une étude comparative entre les modèles de références de l'apprentissage profond développés pour la reconnaissance d'objets tels que VGG19, VGG16, ResNet-152, Inception-V3, Inception-V4, …etc. avec de différents niveaux de fine-tuning et les adapter pour la reconnaissance d'activités humaines. Cependant, cela nécessite de gros investissements matériels, avec une station de travail équipée de cartes graphiques performantes, ou l'utilisation des nouveaux services de calcul dans le cloud tel que AWS de Amazon, Microsoft Azure ou Google Colaboratory.

Nous espérons que ce travail puisse servir utilement d'une approche utile à ceux qui aborderont l'approfondissement de l'étude de la reconnaissance d'activités humaines.

Enfin, le plus important, c'est que le domaine de la reconnaissance d'activités humaines s'impose comme un sujet d'actualité et reste ouvert afin de bénéficier de toute amélioration ou d'idées nouvelles c'est cela qui nous laisse penser que le travail que nous avons réalisé reste ouvert à d'éventuelles améliorations.

Khelalef. A, Ababsa. F, Benoudjit. N, « *A Simple Human Activity Recognition Technique Using DCT*», In: Advanced Concepts for Intelligent Vision Systems. ACIVS 2016. Lecture Notes in Computer Science, vol 10016. pp37-46, Springer, Cham. Lecce, Italy.

Khelalef. A, Benoudjit. N, Abbabsa. F, « *A New Space - Time Technique for Human Activity Recognition using Skeletons and Discrete Cosine Transform*», International Conference on Embedded Systems in Telecommunications and Instrumentation (ICESTI'16), Annaba, Algeria, 2016.

Khelalef. A, Ababsa. F, Benoudjit. N, « *An Efficient Human Activity Recognition Technique Based on Deep Learning*», Pattern recognition and image analysis, Vol. 29, No. 4, pp. 702–715, 2019.

Khelalef. A, Ababsa. F, Benoudjit. N, « A New *Human Activity Recognition technique using YOLO* », Paper submitted to « Pattern recognition and Image Analysis » journal for publication, currently, paper under review.

Références.

[1] C. Harris et M. Stephens, (1988). A Combined Corner and Edge Detector. Dans Alvey Vision Conference, pp 147–151, Manchester, Royaume-Uni.

[2] Lowe D.G. Object Recognition from Local Scale-invariant Features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–25 September 1999; Vol 2, pp. 1150–1157.

[3] H. Bay, T. Tuytelaars, and L. J. Van Gool. Surf: Speeded up robust features. In Proceedings of the European Conference on Computer Vision (ECCV), Graz, Austria, pp 404-417, 2006.

[4] Ojala, T., Pietikainen, M. et Harwood, D. (1996). A comparative study of texture measures with classification based on feature distribution. Pattern Recognition, Vol 29(1) pp. 55-59.

[5] Blank M., Gorelick L., Shechtman E., Irani M. & Basri R. (2005) Actions as Space-Time Shapes. In Proc. ICCV, pp. 1395 – 1402.

[6] Aaron F, James W. “The Recognition of Human Movement Using Temporal Templates”, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 23, no. 3, March 2001.

[7] Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Vol 1, pp. 886–893.

[8] W. Lu; J.J. Little, (2006). Simultaneous tracking and action recognition using the PCA-HOG de-scriptor. In Proceedings of the 3rd Canadian Conference on Computer and Robot Vision, Quebec, PQ, Canada, 7–9 June 2006; pp. 6.

[9] Sheikh, Y., Sheikh, M., and Shah, M. (2005). Exploring the space of a human action. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1 - Vol 01, ICCV ’05, pp 144–149, Washington, DC, USA. IEEE Computer Society.

[10] Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3d joints. In CVPR Workshops, pp 20–27. IEEE.

[11] Fujiyoshi H, Lipton A J, (2004). Real-time human motion analysis by image skeletonization. IEICE Transactions on Information and Systems E Series D, Vol 87(1), pp 113-120.

[12] Sedai S, Bennamoun M and Huynh D, (2009). Context-based Appearance Descriptor for 3D Human Pose estimation from Monocular Images, In Proc. of DICTA, Digital Image Computing: Techniques and Applications, pp 484–491.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F.-F. Li, (2014). Large scale video classification with convolutional neural networks. in Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014) (Columbus, OH), pp. 1725–1732.

[14] T. Dobhal, V. Shitole, G. Thomas, and G. Navada, (2015). Human activity recognition using Binary Motion Image and deep learning. Procedia Comput. Sci. Vol 58, pp 178–185.

[15] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, (2011). Sequential deep learning for human action recognition. In Human Behavior Understanding, Proc. Second International Workshop, HBU 2011, Ed. By A. A. Salah and B. Lepri, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg),Vol. 7065, pp. 29–39.

[16] P. Wang, W. Li, Z. Gao, J. Zhang, T. Chang, and P. O, (2016). Ogunbona. Action recognition from depth maps using deep convolutional neural networks. IEEE Trans. Human-Mach. Syst. Vol 46 (4), pp 498–509 .

[17] P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento, (2014). Exploiting the deep learning paradigm for recognizing human actions," in Proc. 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2014) (Seoul, South Korea, 2014), pp. 93–98.

[18]. K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos, arXiv preprint arXiv:1406.2199. 2014. https://arxiv.org/abs/1406.2199.

[19] Mouna Selmi, (2014). Reconnaissance d'activités humaines à partir de séquences vidéo. Thèse de doctorat, Institut National des Télécommunications, France.

[20] Thomas B. Moeslund, Adrian Hilton and Volker K, (2006). A survey of advances in vision-based human motion capture and analysis. Computer Vision and Image Understanding, Vol. 104, no. 2-3, pp 90–126.

[21] T. Ahmad and J. Rafique, (2015). Using Discrete Cosine Transform Based Features for Human Action Recognition, Journal of Image and Graphics, Vol. 3, No. 2.

[22] Weinland, D., Ronfard, R., and Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. Comput. Vis. Image Underst., Vol 104(2), pp 249–257.

[23] Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing action at a distance. In proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2, ICCV '03, pages 726–, Washington, DC, USA. IEEE Computer Society.

[24] Cyrille Migniot, (2012). Segmentation de personnes dans les images et les vidéos, thèse de doctorat, Université de Grenoble, France.

[25] Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In IN ICCV, pp 432–439.

[26] Laptev, I., Caputo, B., Schüldt, C., and Lindeberg, T. (2007). Local velocityadapted motion events for spatio-temporal recognition. Computer Vision and Image Understanding, Vol 108(3), pp 207–229.

[27] Dollár, P.; Rabaud, V.; Cottrell G.; Belongie, S. Behavior Recognition via Sparse Spatio-Temporal Features. In Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2005; pp. 65–72.

[28] M. Bregonzio, Shaogang Gong and Tao Xiang. Recognising action as clouds of space-time interest points. 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 1948-1955, doi: 10.1109/CVPR.2009.5206779.

[29] Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA '07, pp 357–360, New York, NY, USA. ACM.

[30] P. Liu, J. Wang, M. She and H. Liu. Human action recognition based on 3D SIFT and LDA model. 2011 IEEE Workshop on Robotic Intelligence In Informationally Structured Space, Paris, 2011, pp. 12-17, doi: 10.1109/RIISS.2011.5945790.

[31] Manal A G, Lei Z and Yoshihiko G, (2012). Spatio-temporal SIFT and Its Application to Human Action Classification. ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 301–310.

[32] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, Vol 1, pp 511, Los Alamitos, CA, USA, 2001. IEEE Computer Society.

[33] Willems, G., Tuytelaars, T., and Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. In Proceedings of the 10th European Conference on Computer Vision : Part II, ECCV '08, pp 650–663, Berlin, Heidelberg. Springer-Verlag.

[34] Samir Sahli, (2013). Détection robuste et automatique de véhicules dans les images aériennes. Thèse de doctorat, université LAVAL, Québec, Canada.

[35] Olivier Schwander, (2017). Identification de visages. Rapport de stage, INRIA Rhône-Alpes.

[36] M. Pietikäinen et al., Computer Vision Using Local Binary Patterns, chapter 2, Local Binary Patterns for Still Images, pp 13-47, 2011.

[37] Kellokumpu V., Zhao G. & Pietikäinen M. (2009), Human Activity Recognition Using a Dynamic Texture Based Method, in Proc BMVC, pp 10.

[38] Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. Perception & Psychophysics, 14 :201–211.

[39] Redha Touati, (2014). Reconnaissance des actions humaines à partir d'une séquence vidéo. Mémoire pour l'obtention du grade de maitre en sciences, Université de Montréal.

[40] Mathieu Barnachon, (2013). Reconnaissance d'actions en temps réel à partir d'exemples. Thèse de doctorat, université de Lyon, 2013.

[41] Lin, Z., Jiang, Z., and Davis, L. S. (2009). Recognizing actions by shapemotion prototype trees. In ICCV, pp 444–451. IEEE.

[42] D. Weinland and E. Boyer, (2008). Action recognition using exemplar-based embedding. 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, pp. 1-7, doi: 10.1109/CVPR.2008.4587731.

[43] Vapnik, V. N. 1995, The nature of statistical learning theory (Springer Book).

[44] Imtiaz H et al, (2015). Human Action Recognition based on Spectral Domain Features. 19th Annual Conference, KES-2015, Singapore.

[45] Schuldt, C.; Laptev, I.; Caputo, B, (2004). Recognizing Human Actions, A Local SVM Approach. In Proceedings of the 17th IEEE International Conference on Pattern Recognition (ICPR), Cambridge, UK, Vol 3, pp. 32–36.

[46] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, (2008). Learning realistic human actions from movies. 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, pp. 1-8, doi: 10.1109/CVPR.2008.4587756.

[47] Jhuang, H., T. Serre, L. Wolf et T. Poggio. 2007, A biologically inspired system for action recognition, in Internationale Conference on Computer Vision, pp. 1-8.

[48] H. Foroughi, A. Naseri, A. Saberi and H. Sadoghi Yazdi, (2008). An eigenspace-based approach for human fall detection using Integrated Time Motion Image and Neural Network. 2008 9th International Conference on Signal Processing, Beijing, pp. 1499-1503, doi: 10.1109/ICOSP.2008.4697417..

[49] Fiaz M.K.; Ijaz, B (2010). Vision based Human Activity Tracking using Artificial Neural Networks. In Proceedings of IEEE International Conference on Intelligent and Advanced Systems (ICIAS), Kuala Lumpur, Malaysia, 15–17, pp.1–5.

[50] Sehad Mounir, (2015). Segmentation d'image par une approche basée sur les caractéristiques texturales, temporelles et spectrales : application aux images MSG. Thèse de doctorat, université de Tizi Ouzou, Algérie,

[51] Iheb Ben Amor, (2014). Gestion de la collaboration et compétition dans le crowdsourcing : une approche avec prise en compte de fuites de données via les réseaux sociaux. Thèse de doctorat, université Paris Descartes.

[52] Yamato, J.; Ohya, J.; Ishii, K, (1992). Recognizing Human Action in Time-sequential Images using Hidden Markov Model. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Champaign, IL, USA, 15–18; pp. 379–385.

[53] Weinland, D., E. Boyer et R. Ronfard, (2007). Action recognition from arbitrary views using 3d exemplars. in Internationale Conference on Computer Vision, pp. 1-7.

[54] Ikizler, N. et D. A. Forsyth. 2008, Searching for complex human activities with no visual examples, International Journal of Computer Vision, Vol. 80, no 3, pp. 337-357.

[55] Chakraborty, B., O. Rudovic et J. Gonzalez, (2008). View-invariant humanbody detection with extension to human action recognition using componentwise HMM of body parts. in ICAFGR, pp. 1-6.

[56] Joris Guerry, (2017). Reconnaissance visuelle robuste par réseaux de neurones dans des scénarios d'exploration robotique. Détecte-moi si tu peux ! , thèse de doctorat, Université Paris-Saclay, 2017. France.

[57] Eric Gauthier, (1999). Utilisation des réseaux de neurones artificiels pour la commande d'un véhicule autonome. Thèse de doctorat, Institut National Polytechnique de Grenoble - INPG.

[58] Abdessalem Chamekh, (2008). Optimisation des procédés de mise en forme par les réseaux de neurones artificiels. Thèse de doctorat, Mécanique. Université d'Angers.

[59] Quentin Fresnel (2015). Apprentissage de descripteurs audio par Deep learning, application pour la classification en genre musical. Rapport de stage, IRCAM.

[60] Matthew D. Zeiler and Rob Fergus, (2014). Visualizing and Understanding Convolutional Networks. ECCV 2014, Part I, LNCS 8689, pp. 818–833.

[61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Advances in neural information processing systems Vol 25(2).

[62] Prajit Ramachandran, Barret Zoph, Quoc V. Le. Searching For Activation Functions. https://arxiv.org/abs/1710.05941v2 (Octobre 2020).

[63] Moualek Djaloul Youcef, (2017). Deep Learning pour la classification des Images. Mémoire de fin d'etudes, université de Telemcen, 2017.

[64] http://www.image-net.org/ (Octobre 2020)

[65] http://cocodataset.org/ (Octobre 2020)

[66] http://host.robots.ox.ac.uk/pascal/VOC/ (Octobre 2020)

[67] Alfredo Canziani & Eugenio Culurciello, Adam Paszke, (2017). An Analysis of Deep Neural Network Models for Practical Applications. arXiv:1605.07678.

[68] Yann LeCun, Yoshua Bengio, (1995). convolutional networks for images, speech, and time-series. The handbook of brain theory and neural networks 3361 (10).

[69] Karen Simonyan & Andrew Zisserman, (2015). « Very Deep Convolutional NetworksnFor Large-Scale Image Recognition », ICLR 2015.

[70] J. S. Combinido, J. R. Mendoza and J. Aborot, (2018). A Convolutional Neural Network Approach for Estimating Tropical Cyclone Intensity Using Satellite-based Infrared Images. 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, pp. 1474-1480, doi: 10.1109/ICPR.2018.8545593.

[71] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.

[72] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv :1602.07261.

[73] Sidike, Paheding & Alom, Md. Zahangir & Taha, Tarek & Asari, Vijayan, (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. arXiv:1803.01164.

[74] K. He, X. Zhang, S. Ren, and J. Sun, (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

[75] B. R. Abidi, Y. Zheng, A. V. Gribok, and M. A. Abidi, (2006). Improving weapon detection in single energy X-ray images through pseudocoloring. IEEE Trans. Syst. Man Cybern. C, Appl. Rev., Vol. 36, No. 6, pp. 784–796.

[76] Kumari, S.; Mitra, S.K, (2011). Human Action Recognition Using DFT. In Proceedings of the third IEEE National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Hubli, India, 15–17, pp. 239–242.

[77] T. Ahmad and J. Rafique, (2015). Using Discrete Cosine Transform Based Features for Human Action Recognition. Journal of Image and Graphics, Vol. 3, No. 2.

[78] H. Imtiaz, et al, (2015). Human Action Recognition based on Spectral Domain Features, 19th Annual Conference, KES-2015, Singapore.

[79] Z. Jiang, Z. Lin, and L. Davis, (2012). Recognizing human actions by learning and matching shape-motion prototype trees. IEEE Trans. Pattern Anal. Mach. Intell. 34 (3), 533–547.

[80] R. M. Zhao, H. Lian, H. W. Pang, B.N. Hu, (2008). A Watermarking Algorithm by Modifying AC Coefficies in DCT Domain. International Symposium on Information Science and Engieering, pp. 159-162. IEEE.

[81] N. Otsu, (1979). A threshold selection method from Gray-level histograms. IEEE Trans. Syst., Man, Cybern. Vol 9 (1), pp 62–66.

[82] A. Petros, W. Ronald, (1986). Morphological Skeleton Representation and Coding of Binary Images. IEEE Trans On Acoustic speech and signal processing, Vol ASSP-34, No 5.

[83] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, (2016). You Only Look Once: Unified, Real-Time Object Detection. CVPR 2016, OpenCV People's Choice Award.

[84] https://pjreddie.com/darknet/yolo/ (Octobre 2020)

[85] O. Boiman & M. Irani, (2006). Similarity by Composition. In Proc. Neural Information Processing Systems (NIPS).

[86] L. Wang and D. Suter, (2007). Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model. 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, pp. 1-8, doi: 10.1109/CVPR.2007.383298.

[87] V. Kellokumpu, Zhao G. & Pietikäinen M, (2008). Texture Based Description of Movements for Activity Analysis. In Proc. VISAPP, Vol. 1, pp. 206 – 213

[88] S. Wong and R. Cipolla, (2007). Extracting spatiotemporal interest points using global information. In Proc. 2007 IEEE 11th International Conference on Computer Vision (ICCV) (Rio de Janeiro, Brazil), pp. 1‑8.

[89] J. C. Niebles, H. Wang, and F.-F. Li, (2007). Unsupervised learning of human action categories using spatial-temporal words. Int J. Comput. Vision, Vol 79 (3), pp 299‑318.

[90] B. Chen, J.-A. Ting, B. Marlin, and N. de Freitas, (2010). Deep learning of invariant spatio-temporal features from video. In Deep Learning and Unsupervised Feature Learning Workshop — NIPS 2010 (24th Annual Conference on Neural Information Processing Systems) (Whistler,Canada), pp. 1‑9.

[91] V. Vivek, Z. Naifan, and G.-J. Qi, (2015). Differential recurrent neural networks for action recognition. In Proc. 2015 IEEE International Conference on Computer Vision (ICCV 2015) (Santiago, Chile), pp. 4041‑4049.

[92] L. Sun, K. Jia, T.-H. Chan, Y. Fang, G. Wang, and S. Yan, (2014). DL-SFA: Deeply-learned slow feature analysis for action recognition. in Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2014) (Columbus, OH), pp. 2625‑2632.