Développement d'une méthode d'extraction automatique de métriques de performance en boxe basée sur la vision par ordinateur

par

Juliette SEMINARO

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE EN GÉNIE DES TECHNOLOGIES DE LA SANTÉ
M. Sc. A.

MONTRÉAL, LE 14 SEPTEMBRE 2022

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC


Juliette Seminaro, 2022

Cette licence Creative Commons signifie qu'il est permis de diffuser, d'imprimer ou de sauvegarder sur un autre support une partie ou la totalité de cette œuvre à condition de mentionner l'auteur, que ces utilisations soient faites à des fins non commerciales et que le contenu de l'œuvre n'ait pas été modifié.

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. David Labbé, directeur de mémoire
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Sheldon Andrews, codirecteur de mémoire
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Jocelyn Faubert, codirecteur de mémoire
École d'optométrie à l'Université de Montréal

M. Julien Clément, président du jury
Département de génie des systèmes à l'École de technologie supérieure

M. Carlos Vázquez, membre du jury
Département de génie logiciel et des TI à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 31 AOÛT 2022

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

J’aimerais remercier David Labbé pour son rôle comme directeur de recherche, et toutes les responsabilités impliquées. Merci pour l’opportunité d’entrer dans le monde de la recherche dans un environnement aussi stimulant que le LIO au CRCHUM. Merci de ton accompagnement au cours de ces dernières années, qui n’ont pas été les plus paisibles, ni les plus prévisibles. Tu as été un directeur disponible, patient et compréhensif sur lequel je pouvais compter si je me heurtais à un problème. En ayant travaillé dans ton équipe de recherche, j’en ressors avec davantage de connaissances, de savoir-être et de savoir-faire. En te souhaitant le meilleur pour la suite.

Merci à mon codirecteur de recherche Sheldon Andrews, qui m’a beaucoup appris sur le monde de l’informatique et de l’apprentissage machine. Ton expertise en la matière m’a permis de naviguer plus facilement dans cet univers qui constituait une nouveauté pour moi. Tes commentaires et suggestions toujours pertinents m’ont été de grande valeur.

Merci également à mon second codirecteur Jocelyn Faubert. Bien que le projet se soit quelque peu éloigné des axes de recherche de l’École d’optométrie de l’Université de Montréal, tes interventions toujours éclairantes ont été grandement appréciées.

Un grand merci également à Mathieu Charbonneau et Thomas Romeas pour l’opportunité de travailler sur un projet aussi intéressant, concret et appliqué au sport comme celui-ci. Votre lien direct avec l’Institut national du sport du Québec et ses athlètes, votre expertise ainsi que votre engagement dans ce projet m’ont été précieux.

Finalement, merci à Sara Saint-Pierre Côté pour son aide à la Qualité de ce projet. Merci de ton accompagnement, en plus de ton infinie patience.

Développement d'une méthode d'extraction automatique de métriques de performance en boxe basée sur la vision par ordinateur

Juliette SEMINARO

RÉSUMÉ

En boxe, plusieurs études ont démontré que l'efficience des coups d'un athlète, c'est-à-dire le ratio entre les coups atteignant leur cible sur le total des coups lancés, est un indicateur clé d'une victoire lors d'un combat. Les mouvements des boxeurs et leur contrôle du ring sont également importants puisqu'il s'agit d'éléments considérés dans le système de pointage de la boxe. Il est donc essentiel de faire le suivi de telles métriques afin de mesurer les performances des athlètes ou encore, obtenir de l'information stratégique sur le profil des adversaires. Toutefois, le visionnement des vidéos de combat ainsi que l'annotation des divers événements est une tâche chronophage, principalement effectuée par un expert analyste des performances. Des tentatives d'automatisation de ce processus requièrent de l'équipement spécialisé tel que des caméras à temps de vol, des systèmes de capture de mouvement ou des capteurs inertiels. De plus, il n'est pas pratique pour les athlètes d'être munis de marqueurs réfléchissants ou de capteurs pendant des combats, puisqu'ils peuvent entraver leurs mouvements, être endommagés ou même blesser les boxeurs. Dans le cadre de cette maîtrise, une méthode d'extraction automatique de métriques de performance en boxe basée sur la vision par ordinateur a été développée. Celle-ci permet d'extraire le nombre et le type de coups de poing donnés par un boxeur à partir d'une vidéo de *shadow boxing* (méthode d'entraînement où l'athlète boxe dans le vide en imaginant un adversaire), en plus de générer des courbes et des cartes de chaleur permettant de quantifier le contrôle du ring par les athlètes. En utilisant uniquement des images enregistrées par une seule caméra vidéo, la méthode est donc accessible puisqu'elle est abordable et ne requiert pas d'équipement spécialisé. Elle permet d'obtenir une exactitude de classification moyenne pondérée de 77 % sur les positions et techniques offensives suivantes : garde ouverte, garde, direct avant, direct arrière, crochet avant, crochet arrière, uppercut avant et uppercut arrière. La méthode d'extraction automatique de métriques de performance permet également de faire le suivi des boxeurs sur l'aire de combat à 12,2 cm près.

Mots-clés : Boxe, vision par ordinateur, apprentissage machine, analyse des performances

Development of an Automated Method for Performance Analysis in Boxing based on Computer Vision

Juliette SEMINARO

ABSTRACT

In boxing, studies have shown that punching efficiency, *i.e.* ratio between punches landed and the total number of punches thrown, is key for winning a bout. The movements of the boxers and their control of the ring space are also important as they are part of ring generalship, which is a component that is considered in the boxing scoring system. Therefore, monitoring such metrics is crucial to assess an athlete's performance or to gain knowledge about a boxer's fighting style. However, subjectively observing and manually annotating these events during a bout is a tedious task that must be done by a sport performance analyst. Previous attempts to automate the process required specialized equipment such as time-of-flight cameras, motion capture systems or inertial sensors. The markers and sensors required by these methods are not practical to use in real bouts since they can disrupt the natural motion of the boxers, may be damaged upon contact, or worst, may injure the athletes. In this work, we propose an innovative computer vision-based method to automatically extract performance metrics such as the number and type of punches thrown by a boxer from a shadowboxing video, and ring control through trajectories and heatmaps from a bout video. Our cost-effective approach requires only monocular images recorded by a single video camera, which eliminates the need of any specialized equipment. On average, it achieves 77 % weighted classification accuracy of the following stances and punches: unguarded, guarded, jab, cross, lead hook, rear hook, lead uppercut and rear uppercut. Moreover, it can track a boxer's position in the ring within a 12,2 cm margin of error on average.

Keywords: Boxing, computer vision, machine learning, performance analysis

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AGC-LSTM	*Attention enhanced Graph Convolutional* LSTM
AUC	Aire sous la courbe
Bi-LSTM	LSTM bidirectionnel
C	Direct arrière à la tête
CNN	Réseau de neurones convolutif
CPN	Réseau de neurones de pyramides en cascade
FGI	Inférence gaussienne floue
G	Garde
GCN	Réseau de neurones convolutif de graphe
GC-LSTM	*Graph Convolutional* LSTM
GMM	Modèle de mélange gaussien
HATNet	*Holistic Appearance and Temporal Network*
HRNet	Réseau de neurones haute résolution
IMU	Unité de mesure inertielle
INS Québec	Institut national du sport du Québec
I3D	*Inflated 3D Convolutional Network*
J	Direct avant
KNN	K plus proches voisins
Lr	Taux d'apprentissage (*learning rate*)
LC	Direct arrière au corps
LGD-3D	Réseau de neurones à diffusion locale et globale tridimensionnelle

LH	Crochet gauche à la tête
LLH	Crochet gauche au corps
LSTM	Réseau à longue mémoire court terme
LR	Régression logistique
LSVM	Machine à vecteur de support linéaire
MLP	Perceptron multicouche
MPJPE	Erreur de position moyenne par partie du corps
PARE	*Part Attention Regressor for 3D human body Estimation*
PCK	Pourcentage de parties du corps bien positionnées
P-LSTM	*Part-Aware* LSTM
VA-LSTM	*View-Adaptative* LSTM
ResNet	Réseau de neurones résiduel
RF	Forêt aléatoire
RH	Crochet droit
RNN	Réseau de neurones récurrent
SMPL	*Skinned Multi-Person Linear*
SPIN	*SMPL oPtimization IN the loop*
ST-GCN	*Spatio-Temporal* GCN
SVC	Classificateur à vecteur de support
SVM	Machine à vecteur de support
TCN	Réseau de neurones convolutif temporel

UCF	Université de Floride centrale
U-CondDGConv	Réseau convolutif de graphes directionnels conditionnels en U
VIBE	*Video Inference for Body pose and shape Estimation*
XGB	Amplification extrême du gradient
2D	Bidimensionnel(les)
3D	Tridimensionnel(les)
3DPW	*3D Poses in the Wild*

LISTE DES SYMBOLES ET UNITÉS DE MESURE

cm	centimètre
mm	millimètre
po	pouce
°	degré

INTRODUCTION

Dans le sport de haut niveau, l'analyse vidéo est une pratique incontournable pour les entraîneurs et les athlètes afin de pouvoir évaluer de façon détaillée les performances de ces derniers. Celle-ci permet d'identifier tant les aspects techniques et tactiques gagnants que les faiblesses individuelles des sportifs. En boxe, par exemple, le nombre de coups donnés dans une ronde ainsi que leur efficience sont des indicateurs de performance importants (Ashker *et al*., 2011; Davis *et al*., 2015; Thompson & Lamb, 2016). De plus, l'occupation de l'aire de combat est un élément faisant partie des composantes tactiques évaluées dans le système de pointage officiel en boxe. Il est ainsi essentiel d'effectuer un suivi de telles métriques lors des entraînements et des combats.

Le visionnement du matériel audiovisuel, son annotation et l'élaboration d'une stratégie d'entraînement qui suivent l'analyse vidéo constituent toutefois un processus chronophage devant être mené par un expert analyse des performances. Des tentatives d'automatisation ont été effectuées dans de précédentes études. Dès 2009, Khoury & Liu ont utilisé l'inférence gaussienne combinée à une règle liée au contexte pour reconnaître diverses techniques offensives à partir de données issues d'un système de capture de mouvement Vicon. Pour la même tâche, Kasiri-Bidhendi *et al*. (2015) ont plutôt utilisé des images vues de dessus d'un boxeur et une caméra à temps de vol pour entraîner une machine à vecteur de support (SVM) hiérarchique. D'autres ont entraîné des classificateurs avec les signaux de capteurs inertiels (Soekarjo *et al*., 2019; Worsey *et al*., 2020; Khasanshin, 2021). Ces études obtiennent de bons résultats de classification, allant de 86,7 % à 98 %. Toutefois, ces méthodes nécessitent du matériel supplémentaire comme des caméras infrarouges ou à temps de vol à installer respectivement autour ou au-dessus de l'aire de combat, ou encore des marqueurs réfléchissants ou des capteurs inertiels dont doivent être équipés les athlètes. Ces méthodes sont ainsi peu pratiques pour l'analyse de réels combats, puisque, d'une part, il n'est pas possible de garantir que les boxeurs puissent être munis de capteurs lorsqu'ils se livrent de véritables coups. Cela peut être dangereux pour les athlètes ou nuire à leur liberté de

mouvement. De plus, l'équipement porté par les boxeurs tels que les marqueurs ou les capteurs inertiels sont à risque de bris. D'autre part, l'environnement ne peut pas nécessairement être modifié afin de permettre ces acquisitions de données. Par exemple, les équipes souhaitant utiliser la méthode de Kasiri-Bidhendi *et al*. (2015) doivent fixer une caméra à temps de vol au plafond afin de détenir des images vues de dessus en plus de données de profondeur. Au mieux, elles peuvent s'assurer de satisfaire cette exigence matérielle dans leurs propres installations sportives, mais il est peu probable qu'elles puissent faire modifier la configuration d'autres centres. Il serait donc impossible d'analyser les compétitions externes.

Il est donc pertinent de tirer profit de la vision par ordinateur pour analyser automatiquement des vidéos sportives. Au lieu du long processus d'analyse vidéo mentionné plus haut, il ne suffit ainsi que de fournir des images monoculaires à un réseau de neurones qui en retirera par lui-même les métriques de performance. L'intérêt d'utiliser la vision par ordinateur est grand, puisqu'il est aisé et abordable d'enregistrer des vidéos grâce à une simple caméra. En ne nécessitant que des vidéos comme données d'entrée, le besoin d'utiliser divers marqueurs et capteurs est écarté. Ceux-ci sont susceptibles aux bris lors de vrais combats, et peuvent entraver les mouvements des athlètes. De plus, il est ainsi possible d'analyser les vidéos de combat autant en entraînement, en compétition à domicile ou à l'externe. Dans le cadre de cette maîtrise, une méthode d'extraction automatique de métriques de performance en boxe est ainsi développée. Basée sur la vision par ordinateur et l'apprentissage machine, elle permet dans un premier temps de quantifier les différentes techniques offensives effectuées par le boxeur, ainsi que d'identifier son état de vulnérabilité. La reconnaissance d'actions est effectuée sur des vidéos de *shadow boxing* sur lesquelles ne figure qu'un athlète. Ceci permet d'éviter le problème d'occlusion entre deux personnes, et assure ainsi la qualité des données d'entrée. De plus, un suivi des déplacements des athlètes sur l'aire de combat est effectué afin de pouvoir analyser l'occupation de la surface pendant la joute. Contrairement à la reconnaissance d'actions, cette portion de l'outil ne se limite pas à l'analyse de vidéos de *shadow boxing* et permet de traiter de réels combats entre deux adversaires. Bref, deux

contributions dans le domaine de l'analyse de la performance en boxe par la vision par ordinateur sont apportées par ce projet de maîtrise, soit :

1) Un module de reconnaissance d'actions spécifiques à la boxe exact à 77 %;
2) Un algorithme de suivi exact à 12,2 cm près pour la quantification du mouvement et de l'occupation de l'aire de combat des boxeurs.

Le développement de l'outil d'analyse automatique des performances en boxe est effectué en collaboration avec l'Institut national du sport du Québec (INS Québec), qui est à la fois un centre d'entraînement pour les boxeurs d'élite, et un lieu de recherche et développement de techniques et de technologies ayant pour but d'optimiser les performances des athlètes. Cette collaboration avec les experts et les athlètes de l'INS Québec permet de détenir de précieuses informations sur les besoins technologiques du domaine, en plus de multiples données d'entraînements utiles au développement et à l'évaluation de la méthode d'extraction automatique de métriques de performance en boxe. Celle-ci leur permettra en retour d'obtenir facilement lesdites métriques, dont la collecte est actuellement coûteuse en temps et en personnel qualifié.

Dans un premier temps, ce mémoire présente une revue de la littérature permettant d'obtenir une vue d'ensemble sur l'analyse de performance en boxe, de même que sur les notions essentielles de la vision par ordinateur et de la reconnaissance d'actions. Puis, la problématique et les objectifs liés au projet sont détaillés. La méthode automatisée d'extraction de métriques de performance est par la suite présentée de manière approfondie, suivie des résultats quantitatifs et qualitatifs obtenus. Finalement, les points méthodologiques auxquels il faut porter attention, l'intérêt d'une telle méthode et les défis subsistants sont discutés.

CHAPITRE 1

REVUE DE LA LITTÉRATURE

1.1 Analyse des performances en boxe

L'objectif de la présente maîtrise étant d'extraire automatiquement des métriques de performance en boxe, cette section présente dans un premier temps les définitions des techniques d'intérêt dans ce sport. Puis, une revue de la littérature en ce qui a trait aux métriques significatives en boxe et aux méthodes d'extraction automatique de celles-ci est effectuée.

1.1.1 Définitions

Cette sous-section présente au lecteur les définitions de diverses techniques en boxe qui sont mentionnées tout au long de ce mémoire. Il y a six principales techniques offensives en boxe qui peuvent être décrites selon si elles sont dirigées à la tête ou au corps. Ainsi, il y a le direct avant (*jab*), le direct arrière (*cross*), le crochet avant (*lead hook*), le crochet arrière (*rear hook*), l'uppercut avant (*lead uppercut*) et l'uppercut arrière (*rear uppercut*).

Le direct avant consiste en l'extension rapide du coude du même côté que le pied avant (Figure 1.1). Le bras est idéalement parallèle au sol pour une plus grande puissance effective, et la paume de la main fait face à ce dernier. Il s'agit de la technique offensive la plus utilisée en boxe, puisqu'elle peut être exécutée rapidement tout en conservant une certaine position de garde. Elle permet également de tenir l'adversaire à distance (Werner & Lachica, 2000).

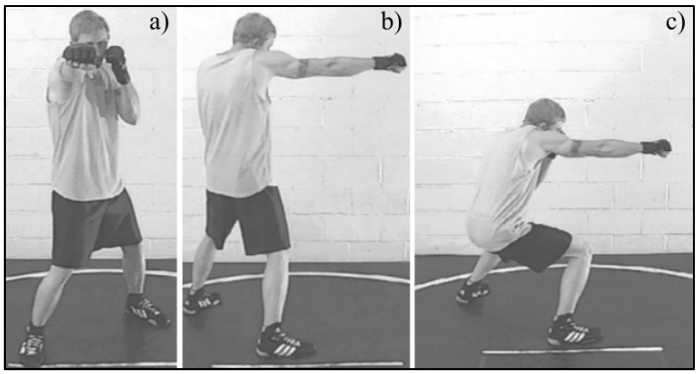


Figure 1.1 Direct avant à la tête (a et b) et au corps (c)
Adaptée de Hatmaker & Werner (2004)

Le direct arrière consiste aussi en l'extension du coude, le bras parallèle au sol, mais celui-ci est du côté opposé au pied avant (Figure 1.2). Il s'agit d'une technique offensive plus puissante que le direct avant puisqu'elle tire avantage de l'énergie engendrée par la rotation du torse et du pied arrière. En contrepartie, la position du boxeur est instable à la fin de cette rotation, ce qui le rend plus vulnérable jusqu'à ce qu'il revienne en position de garde (Werner & Lachica, 2000).

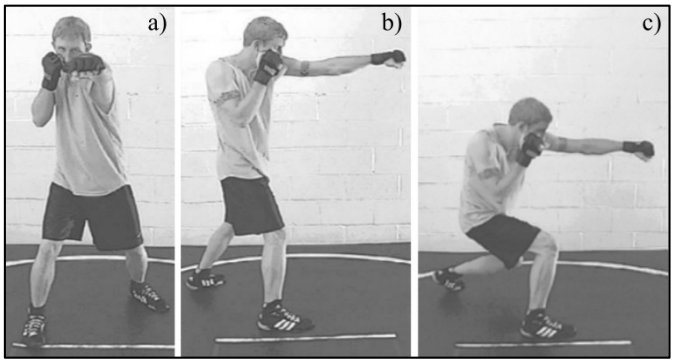


Figure 1.2 Direct arrière à la tête (a et b) et au corps (c)
Adaptée de Hatmaker & Werner (2004)

Lors d'un crochet avant, le coude du bras du même côté que le pied avant est plié à un angle de 90° (Figure 1.3). De manière semblable au direct arrière, la puissance de cette technique offensive provient de la rotation du torse et du pied avant. Ce coup est utilisé lorsque l'adversaire se trouve à une courte distance.

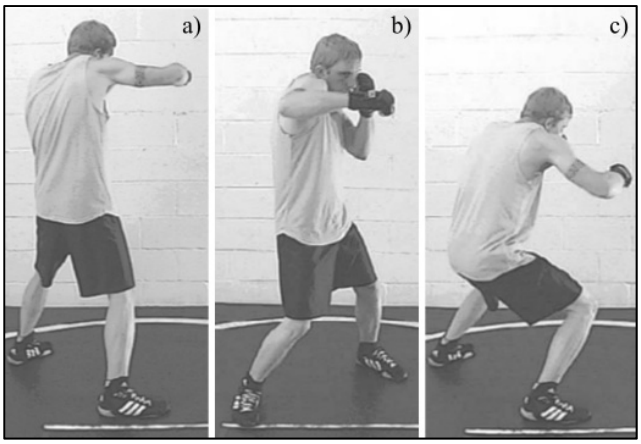


Figure 1.3 Crochet avant à la tête (a et b) et au corps (c)
Adaptée de Hatmaker & Werner (2004)

Le même principe s'applique pour le crochet arrière, mais le bras donnant le coup est celui opposé au pied avant (Figure 1.4). Cette technique offensive, d'autant plus lorsque dirigée vers le corps, est toutefois plus lente et plus facile à lire. Hatmaker & Werner (2004) recommandent ainsi son utilisation en fin de combinaison, ou lorsque la situation est absolument sécuritaire pour son exécution.

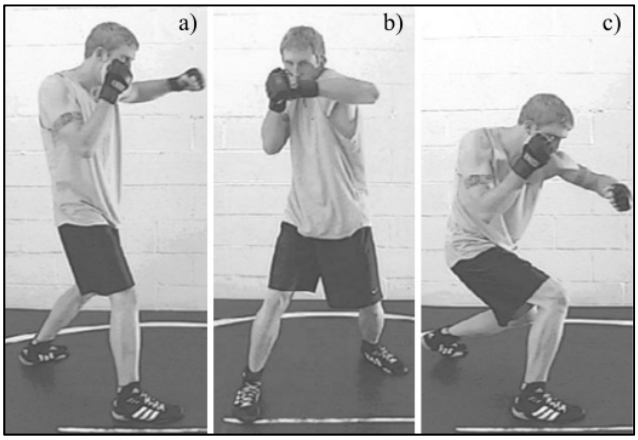


Figure 1.4 Crochet arrière à la tête (a et b) et au corps (c)
Adaptée de Hatmaker & Werner (2004)

Lors d'un uppercut avant, le poing du même côté que le pied avant est descendu de six à huit pouces, puis est remonté rapidement vers le haut afin de livrer le coup (Figure 1.5). Les

cibles sont la mâchoire et le plexus solaire, respectivement pour l'uppercut donné à la tête et l'uppercut donné au corps. Tout comme le crochet, l'uppercut est utilisé lorsque l'opposant est situé à une courte distance. Cette technique offensive est utile pour passer sous la garde adverse, ce qui peut éventuellement créer des ouvertures pour d'autres types de coups. Hatmaker & Werner (2014) recommandent de ne pas commencer une combinaison avec un uppercut, mais plutôt de l'utiliser après un direct avant.

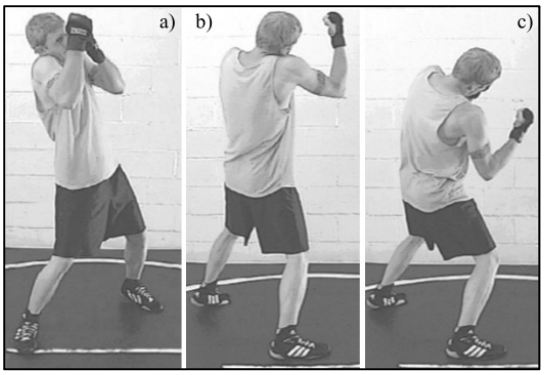


Figure 1.5 Uppercut avant à la tête (a et b) et au corps (c)
Adaptée de Hatmaker & Werner (2004)

Les mêmes considérations s'appliquent pour l'uppercut arrière, exception faite que le bras qui donne le coup est celui opposé à la jambe avant du boxeur (Figure 1.6).

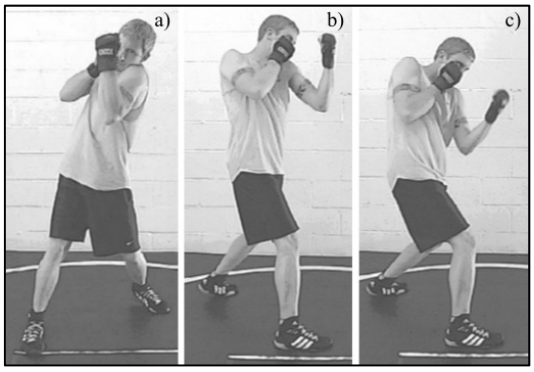


Figure 1.6 Uppercut arrière à la tête (a et b) et au corps (c)
Adaptée de Hatmaker & Werner (2004)

L'état de vulnérabilité faisant partie des composantes que la méthode d'extraction automatique de métriques de performance en boxe doit mesurer, celle-ci doit pouvoir reconnaître si le boxeur se trouve ou non en position de garde (Figure 1.7).

Figure 1.7 Principales positions de garde
Adaptée de Hatmaker & Werner (2004)

1.1.2 Identification des métriques de performance

Plusieurs études ont souhaité identifier les variables déterminantes d'une victoire lors d'un combat de boxe. Ainsi, Ashker (2011) a demandé à cinq arbitres fédérés de visionner les vidéos de 33 combats et d'analyser les techniques offensives et défensives des boxeurs. L'auteur a par la suite conduit des analyses statistiques, ce qui lui a permis de constater que les gagnants lancent davantage de coups, avec de surcroît un plus haut taux de succès. Le nombre de coups donnés et leur efficience comme indicateurs d'une victoire sont également relevés dans les études de Davis *et al*. (2015) et de Thomson & Lamb (2016). Dans l'étude de Davis *et al*. (2015), ce sont les vidéos de 29 combats qui ont été observées par un entraîneur qualifié en boxe pour obtenir manuellement les différentes données d'intérêt, telles

que le temps de chaque ronde de combat, le nombre et le type de coups portés, ainsi que les techniques défensives effectuées. Dans l'étude de Thomson & Lamb (2016), l'auteur principal s'est servi du logiciel Dartfish TeamPro pour faciliter l'annotation manuelle des événements d'intérêt survenant sur les vidéos de 46 combats.

Les conclusions de ces études concordent avec le système de pointage de l'Association Internationale de Boxe. Trois critères à évaluer sont ainsi énoncés dans la réglementation officielle :

1) Le nombre de coups de qualité portés à la cible;
2) La domination du combat par supériorité technique et tactique;
3) La compétitivité.

Outre le nombre de coups donnés et leur efficience, des éléments stratégiques comme l'occupation de l'aire de combat sont également considérés lors de l'évaluation de la performance des boxeurs par les juges. Par exemple, un athlète étant en mesure de contrôler le centre du ring et ainsi de coincer son adversaire entre lui-même et les cordes est généralement vu comme étant en contrôle de l'action, ce qui lui permet d'obtenir un plus haut pointage dans le second et le troisième critère. Les déplacements et l'occupation du ring par les boxeurs sont donc des métriques tout aussi importantes à observer que le nombre de coups et leur efficience pour obtenir un tableau complet des performances de chaque athlète.

1.1.3 Extraction automatique des métriques de performance

L'étape de visionnement et d'annotation des vidéos de combat est chronophage et doit être effectuée par un expert analyse des performances. Des études se sont donc intéressées à automatiser ce processus grâce à différentes méthodes d'apprentissage machine telles que les forêts aléatoires, les machines à vecteurs de support et l'algorithme des k plus proches voisins (KNN). D'autres ont plutôt employé un perceptron multicouche (MLP), soit le plus simple modèle d'apprentissage profond.

Dès 2009, Khoury & Liu proposent une méthode d'apprentissage machine pour reconnaître à partir des données recueillies par un système de capture de mouvement Vicon sept postures en boxe, soit la garde (G), le direct avant (J), le direct arrière à la tête (C), le direct arrière au corps (LC), le crochet droit (RH), le crochet gauche à la tête (LH) et le crochet gauche au corps (LLH). Trois boxeurs de niveau national munis de marqueurs réfléchissants (Figure 1.8) ont ainsi exécuté quatre fois une séquence de 21 coups séparés par une position de garde, résultant en l'acquisition de 252 exemples de techniques offensives. Les auteurs comparent l'inférence gaussienne floue (FGI) couplée à une règle liée au contexte à l'algorithme plus classique du modèle de mélange gaussien (GMM). Ils démontrent ainsi que la FGI présente une performance supérieure au GMM, avec une exactitude de classification moyenne de 87,71 % contre 61 % (Figure 1.9).

Figure 1.8 Environnement d'acquisition des données
Tirée de Khoury & Liu (2009)

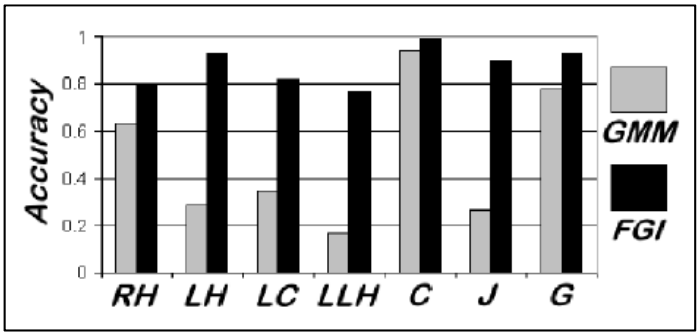


Figure 1.9 Exactitude du FGI
Tirée de Khoury & Liu (2009)

Kasiri-Bidhendi *et al*. (2015) se sont servi d'images vues de dessus (Figure 1.10) ainsi que de données de profondeur capturées par une caméra à temps de vol (Figure 1.11) afin de distinguer six types de coups en boxe, soit le direct avant, le direct arrière, le crochet avant, le crochet arrière, l'uppercut avant et l'uppercut arrière. Huit boxeurs élite ont participé aux séances d'acquisition de données.

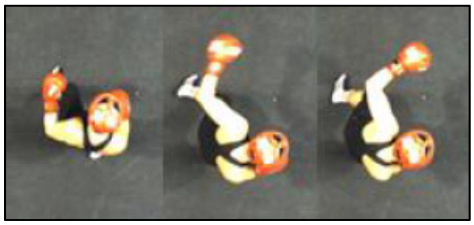


Figure 1.10 Images vues de haut
Tirée de Kasiri-Bidhendi *et al*. (2015)

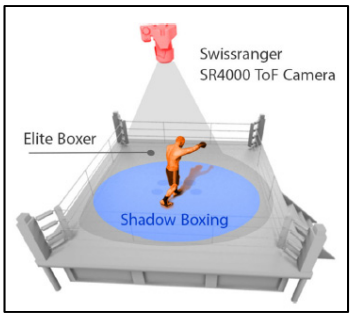


Figure 1.11 Environnement pour la collecte de données
Tirée de Kasiri *et al*. (2017)

Les auteurs entraînent eux aussi des méthodes d'apprentissage machine, soit une machine à vecteur de support et une forêt aléatoire, grâce à 32 exemples de chaque coup, et les évaluent par la validation croisée *Leave-one-out*. Ils testent différentes caractéristiques en entrée des différents modèles, soit NH_{path} la distance entre le cou et la main, h_{max} la distance perpendiculaire entre le coude et la droite reliant le cou et la main, ainsi que les coordonnées sphériques et cylindriques de la main. Ils obtiennent ainsi, avec la SVM hiérarchique, une exactitude de classification supérieure à 90 % (Tableau 1.1 et Figure 1.12).

Tableau 1.1 Exactitude des modèles entraînés par Kasiri-Bidhendi *et al.* (2015)
Tiré de Kasiri-Bidhendi *et al*. (2015)

Caractéristiques d'entrée	**Exactitude (%) selon le modèle**		
	SVM linéaire	**Forêt aléatoire**	**SVM hiérarchique**
Sf1 (NH_{path} + coordonnées sphériques de la main)	85	84	**92**
Cf1 (NH_{path} + coordonnées cylindriques de la main)	86	86	**93**
Sf2 (h_{max} + Sf1)	92	86	**94**
Cf2 (h_{max} + Cf1)	91	88	**96**

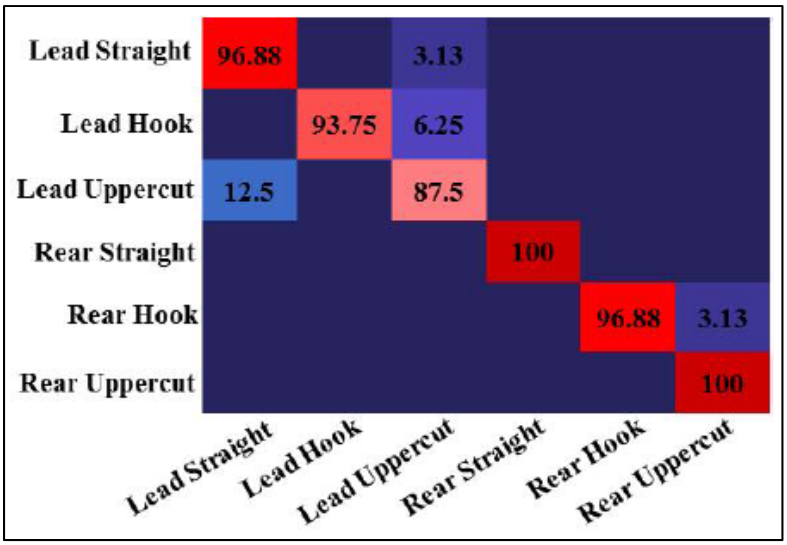


Figure 1.12 Matrice de confusion de la SVM hiérarchique
Tirée de Kasiri-Bidhendi *et al*. (2015)

Soekarjo *et al*. (2019) utilisent un système de capture de mouvement Optotrak et des unités de mesure inertielle (IMU) pour extraire les trajectoires du membre du corps donnant un coup de poing ou un coup de pied. À l'aide de 3 962 trajectoires provenant de 21 participants pratiquant le *kickboxing* et 21 participants ne pratiquant pas de sport de combat, ils entraînent un algorithme des k plus proches voisins et un classificateur à vecteur de support (SVC) à distinguer 12 classes d'actions, réparties sous six techniques de coup de poing et six techniques de coup de pied. Les catégories de coups de poing sont le crochet, le direct, l'uppercut, le coup de poing descendant, le coup de poing Superman et le coup de poing retourné (ces trois derniers coups, qui ne sont pas mentionnés dans les autres études de cette présente revue de la littérature, sont décrits à l'ANNEXE I). Ils obtiennent ainsi une exactitude de classification d'environ 86 %, et ce, pour les deux types de modèles entraînés (Tableau 1.2).

Tableau 1.2 Exactitude des modèles entraînés par Soekarjo *et al*. (2019)
Adapté de Soekarjo *et al*. (2019)

Modèle	Exactitude de reconnaissance du membre donnant le coup (%)	Exactitude de reconnaissance de la technique offensive (%)
KNN	98,9	**86,0**
SVC	99,4	**86,7**

Worsey *et al*. (2020) utilisent les signaux provenant de capteurs inertiels (Figure 1.13) afin d'entraîner différents modèles d'apprentissage machine à reconnaître cinq types de coup de poing, soit le direct avant, le direct arrière, le crochet avant, l'uppercut avant et l'uppercut arrière.

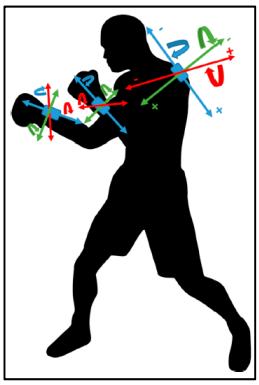


Figure 1.13 Emplacement des capteurs inertiels de Worsey *et al*. (2020)
Tirée de Worsey *et al*. (2020)

Ils comparent ainsi un modèle de régression logistique (LR), une machine à vecteur de support linéaire (LSVM), une machine à vecteur de support gaussienne RBF (GSVM), un perceptron multicouche (MLP), une forêt aléatoire (RF) et un algorithme d'amplification extrême du gradient (XGB). Ils évaluent également deux configurations des capteurs inertiels, où la première ne possède que deux capteurs situés dans les gants du boxeur, tandis que la seconde en ajoute un au niveau de la troisième vertèbre thoracique. Grâce à leur ensemble d'entraînement de 250 coups effectués par un participant pratiquant le muay-thaï,

leur plus haute exactitude de classification est de 98 %, obtenue avec le MLP et la configuration #2 des capteurs.

Tableau 1.3 Exactitude des modèles entraînés par Worsey *et al*. (2020)
Adapté de Worsey *et al*. (2020)

Modèle	**Exactitude (%)**	
	Configuration #1	**Configuration #2**
LR	96	89
LSVM	95	88
GSVM	96	89
MLP	90	**98**
RF	87	83
XGB	79	79

Dans une étude similaire, Khasanshin (2021) entraîne également un MLP à l'aide de données provenant d'IMU (Figure 1.14), soit l'accélération absolue et la vitesse angulaire. L'auteur possède toutefois davantage de données d'entraînement, avec un total de 360 000 coups de poing comprenant les directs, les crochets et les uppercuts. Ceux-ci sont effectués par 85 boxeurs regroupés en trois groupes selon leur niveau d'expertise en boxe.

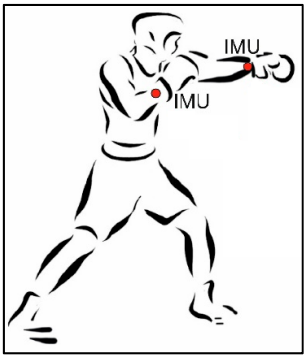


Figure 1.14 Emplacement des IMU de Khasanshin (2021)
Adaptée de Khasanshin (2021)

Il entraîne des modèles spécifiques avec les données exclusives de chaque groupe, en plus d'un modèle universel entraîné avec l'entièreté des données, qu'il évalue par la suite sur

chacun des niveaux d'expertise. Le tableau 1.4 présente les exactitudes de classification de ces modèles.

Tableau 1.4 Exactitude des modèles entraînés par Khasanshin (2021)

Groupe	Exactitude (%)	
	Modèle spécifique	Modèle universel
1 an d'expertise	87,20	91,89
2 à 3 ans d'expertise	**95,33**	**94,17**
5 ans d'expertise	91,73	92,93

Le tableau 1.5 présente un résumé des méthodes présentées dans cette section de la revue de la littérature. Worsey *et al*. (2019) ont la meilleure exactitude de classification, avec 98 %. Toutefois, ils entraînent et évaluent leur modèle avec les données d'un seul participant, ce qui fait qu'il n'y a aucune variabilité inter-sujet en jeu. Kasiri-Bidhendi *et al*. (2015) et Khasanshin (2021) suivent ensuite dans cet ordre, avec des exactitudes respectives rapportées de 96 % et 95,33 %. Dans la première étude, les auteurs détiennent des données d'entrée provenant de deux sources différentes, tandis que la seconde est l'étude disposant du plus grand nombre d'exemples. Finalement, sous les 90 % se trouvent Khoury & Liu (2009) et Soekarjo *et al*. (2019). Ces performances inférieures peuvent s'expliquer dans le premier cas par le fait qu'il s'agit de la seule étude ne recourant pas à l'utilisation des réseaux de neurones. Dans le second cas, les actions d'intérêt sont plus nombreuses, en plus de comporter à la fois des coups de poing et des coups de pied. Il est à noter que ces deux études ont en commun le fait d'utiliser en entrée des données provenant de systèmes de capture de mouvement.

Des méthodes de reconnaissance d'actions en boxe ont ainsi déjà été explorées. Toutefois, elles nécessitent du matériel spécialisé comme un système de capture de mouvement, des caméras à temps de vol, ou encore, des capteurs inertiels que doivent porter les athlètes. Ceci peut constituer un frein à l'implémentation de ces méthodes pour l'analyse de réelles joutes, puisqu'il peut ne pas y avoir ce type de caméra au-dessus de l'aire de combat, et qu'il peut

être impossible pour les boxeurs d'être équipés de capteurs ou de marqueurs réfléchissants durant les duels. Pour éviter ces limitations, l'option de tirer profit des avancées en estimation de pose 3D pour entraîner un classificateur est choisie dans le cadre de cette maîtrise. Les estimateurs de pose 3D sont en fait des réseaux de neurones entraînés à déterminer les coordonnées tridimensionnelles des différentes parties du corps d'une ou plusieurs personnes sur des images. La prochaine section introduit des notions d'apprentissage profond pour mieux présenter les estimateurs de pose à la section 1.3.

Tableau 1.5 Tableau récapitulatif des classificateurs

Auteurs	Meilleur modèle	Type de données d'entrée	Nombre d'exemples	Nombre de sujets	Actions reconnues	Meilleure exactitude (%)
Khoury & Liu (2019)	Inférence gaussienne floue (FGI)	Données tridimensionnelles Vicon (angles d'Euler)	252	3	7 actions : - garde - direct avant - direct arrière à la tête - direct arrière au corps - crochet droit - crochet gauche à la tête - crochet gauche au corps	87,71
Kasiri-Bidhendi *et al*. (2015)	Machine à vecteur de support (SVM) hiérarchique	Images vues de dessus + données de profondeur enregistrées par une caméra à temps de vol	192 (32/coup)	8	6 actions : - direct avant - direct arrière - crochet avant - crochet arrière - uppercut avant - uppercut arrière	96
Soekarjo *et al*. (2019)	Classificateur à vecteur de support (SVC)	Données tridimensionnelles Optotrak (trajectoires) + IMU (segmentation des coups)	3 962	42	12 actions : - crochet - direct - uppercut - coup de poing descendant -coup de poing Superman - coup de poing retourné - coup de pied circulaire - coup de pied avant - coup de pied écrasant - coup de pied latéral - coup de pied arrière - coup de pied tournant	86,7
Worsey *et al*. (2020)	Perceptron multicouche (MLP)	IMU (2-3) (segmentation des coups, accélérations linéaires, accélérations angulaires, rotations axiales)	250	1	5 actions : - direct avant - direct arrière - crochet avant - uppercut avant - uppercut arrière	98
Khasanshin (2021)	Perceptron multicouche (MLP)	IMU (2) (segmentation des coups, accélérations linéaires, accélérations angulaires, rotations axiales)	360 000	85	3 actions : - direct - crochet - uppercut	95,33

1.2 Apprentissage profond et réseaux de neurones

Depuis le MLP, les réseaux de neurones se sont complexifiés pour être aptes à traiter plus de données et à effectuer des tâches plus lourdes en opérations de calcul. Cette section introduit les principaux types de modèles utilisés en estimation de pose 3D et en reconnaissance d'actions.

1.2.1 Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (CNN) portent leur nom de l'opération de convolution effectuée par les couches convolutives du réseau. Celle-ci permet à la fois de réduire la dimensionnalité des données et d'en extraire des informations utiles comme les gradients dans les images.

Puis, Kipf *et al*. (2016) introduisent les réseaux convolutifs de graphe (GCN). Ce type de réseau est capable de réaliser l'opération de convolution sur des graphes, qui sont eux utilisés pour modéliser les diverses relations entre les données par des nœuds et des segments. Les GCN peuvent donc apprendre de la structure même des données.

1.2.2 Réseaux de neurones récurrents

Les réseaux de neurones récurrents (RNN), introduits par Rumelhart *et al*. (1987) sont une famille de réseaux utilisés particulièrement pour la classification ou la prédiction de séries temporelles. Ils possèdent en effet une mémoire interne, ce qui leur permet de retenir l'information de données vues précédemment afin d'obtenir une meilleure capacité de prédiction. Les RNN sont toutefois sujets à la disparition du gradient lorsque la rétropropagation entraîne une multiplication de petites valeurs. Les poids du réseau ne sont donc pas mis à jour, et ce dernier n'apprend plus des données.

Pour pallier ce problème, Hochreiter & Schmidhuber (1997) proposent un réseau à longue mémoire court terme (LSTM), faisant partie de la famille des réseaux de neurones récurrents. Leur unité possède une porte d'oubli, qui permet d'ignorer certains éléments lors de l'apprentissage et ainsi, d'assurer que le gradient ne disparaisse pas durant la rétropropagation (Figure 1.15). D'un point de vue mathématique, une opération d'addition est ajoutée afin d'éviter la multiplication successive de petites valeurs lors de la rétropropagation.

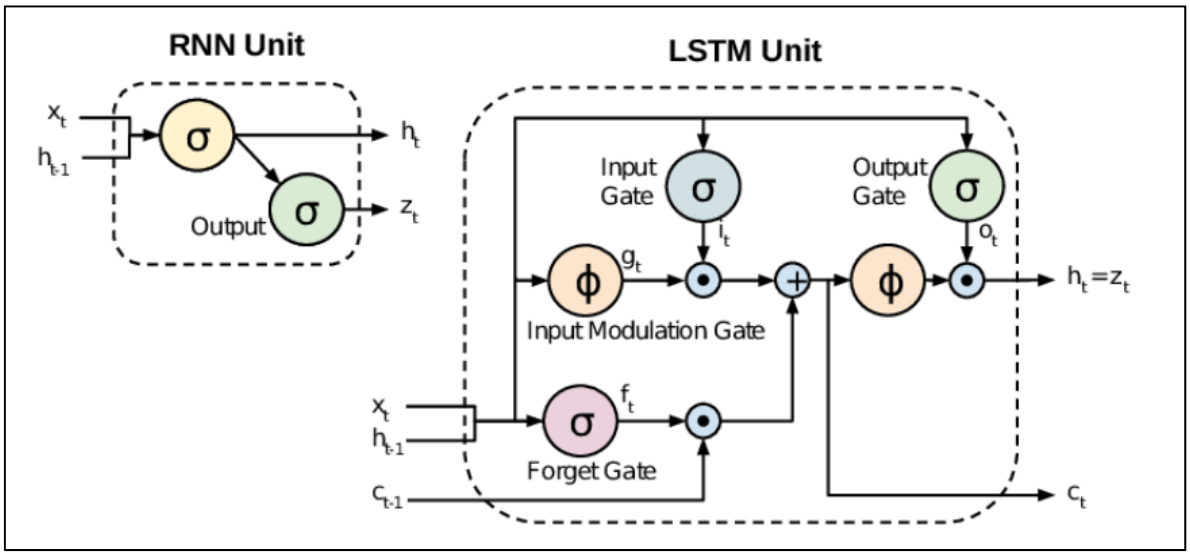


Figure 1.15 Unité RNN comparée à une unité LSTM
Tirée de Tripathi (2021)

1.2.3 Réseaux de neurones transformeurs

À l'heure de la rédaction de ce mémoire, le plus récent type de réseau de neurones est le transformeur (Vaswani *et al*., 2017). Celui-ci se démarque par la présence de modules d'attention dans son architecture (Figure 1.16). Il permet de pallier deux problématiques des RNN, soit le fait que plus la séquence temporelle est longue, plus il est difficile pour le réseau d'apprendre des informations reliant les derniers éléments aux premiers éléments de la série, et le long temps d'entraînement. Le module d'attention permet au réseau d'apprendre quelles parties de la séquence temporelle sont les plus importantes pour la prédiction ou la

classification. Le transformeur se prête bien au calcul parallèle puisqu'il est possible de tronquer la séquence et de donner simultanément ces sous-séries en entrée au réseau. Le temps d'entraînement en est donc réduit.

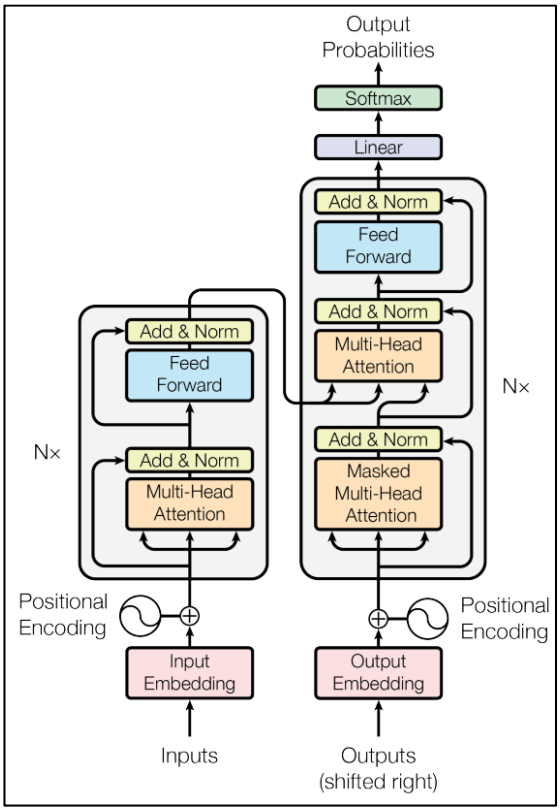


Figure 1.16 Architecture d'un réseau transformeur
Tirée de Vaswani *et al*. (2017)

1.3 Estimation de pose 3D

L'estimation de pose 3D consiste, grâce à un réseau de neurones préentraîné, à obtenir les coordonnées tridimensionnelles d'un objet ou d'une personne à partir d'une image (Figure 1.17).

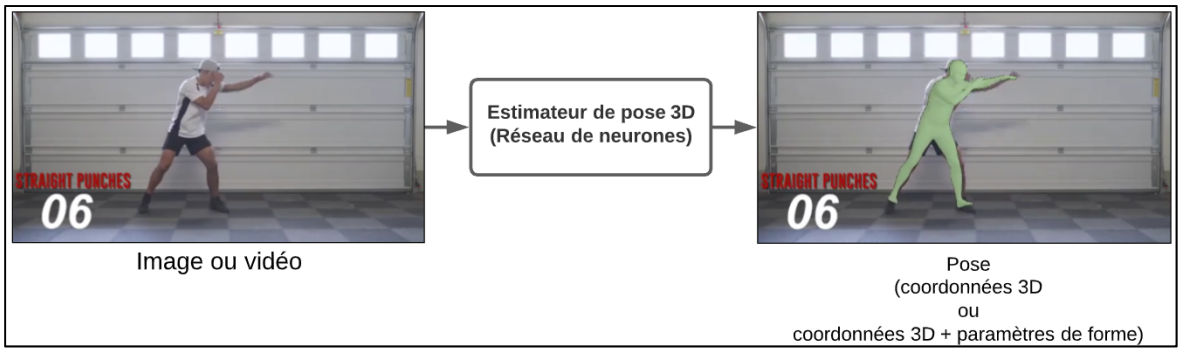


Figure 1.17 Schéma simplifié du processus d'estimation de pose 3D

L'estimation de pose 3D permet ainsi d'obtenir des résultats de même nature qu'un système de capture de mouvement traditionnel, tout en évitant les inconvénients de ce dernier. L'estimation de pose 3D ne requiert que la capture d'images à partir d'une simple caméra, qui sont par la suite données en entrée à un réseau de neurones pour obtenir, en quelques secondes, les coordonnées 3D de la personne figurant sur celles-ci.

Les estimateurs de pose 3D peuvent produire des résultats sous forme de pose, ou encore de pose + forme. La pose est en fait les coordonnées tridimensionnelles de points anatomiques du corps humain. Ces points peuvent être tracés dans un système de coordonnées, en plus des segments les reliant, afin d'obtenir une représentation en bonhomme-allumette de la pose humaine (Figure 1.18). En 2015, Loper *et al*. présentent SMPL, un modèle tridimensionnel qui permet aux estimateurs de pose 3D de produire non seulement la pose, mais aussi la forme humaine (Figure 1.19).

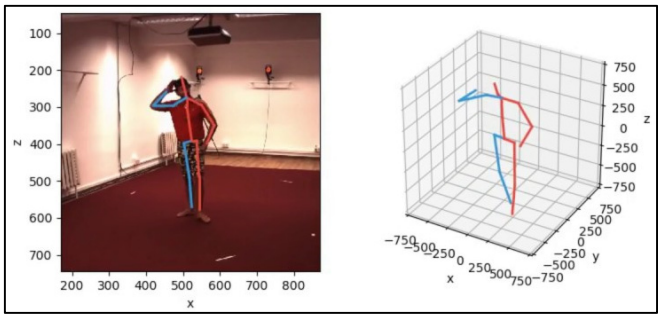


Figure 1.18 Exemple de sortie « pose » d'un estimateur de pose 3D
Tirée de Martinez (2017)

Figure 1.19 Exemple de sortie « pose + forme » d'un estimateur de pose 3D
Tirée de Kolotouros *et al*. (2019)

Dans la littérature, les méthodes d'estimation de pose 3D sont généralement classées en deux types d'approche, soit l'approche par régression et l'approche par optimisation.

1.3.1 Approche par régression (ou bout-à-bout)

Dans l'approche par régression, l'estimateur de pose 3D apprend directement à régresser les coordonnées tridimensionnelles à partir des pixels de l'image (Figure 1.20).

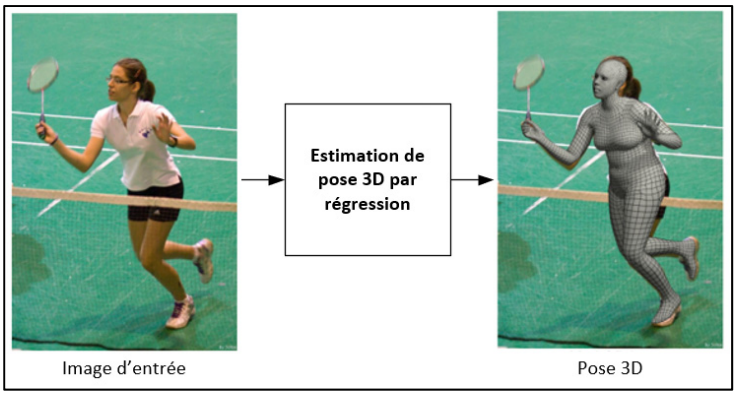


Figure 1.20 Processus d'estimation de pose 3D par régression
Adaptée de Godwisdom (2020)

Pour entraîner un tel modèle, il faut cependant recourir à de larges bases de données ayant une vérité terrain tridimensionnelle. En 2010, Sigal *et al*. ont constitué l'ensemble HumanEva grâce à 40 000 images synchronisées avec des données issues d'un système de

capture de mouvement Vicon. Il s'agissait du premier standard de référence pour évaluer les estimateurs de pose 3D. En 2014, Ionescu *et al*. ajoutent aux ressources disponibles Human3.6M, qui contient plus de 3 600 000 images couplées à des poses tridimensionnelles provenant encore une fois d'un système de capture de mouvement Vicon. Ces précédents ensembles de données sont toutefois enregistrés dans des environnements intérieurs contrôlés. Pour remédier au manque d'images naturelles, Mehta *et al*. (2017) proposent MPI-INF-3DHP, qui contient plus de 1 300 000 images combinées à la vérité terrain 3D. Le fond vert en arrière-plan des images peut être modifié afin de simuler des environnements naturels. Puis, en 2018, von Marcard *et al*. élaborent 3DPW grâce à des capteurs inertiels afin d'obtenir plus de 51 000 images enregistrées directement en environnement naturel ainsi que les poses 3D associées.

Puisque l'estimateur de pose 3D doit traiter une image en entrée, les modèles suivant l'approche par régression sont pour la plupart de type convolutif. Pour estimer la pose, Zhou *et al*. (2016) ajoutent à leur réseau une couche cinématique lui permettant d'apprendre les relations spatiales entre les différentes parties du corps. Pavlakos *et al*. (2017) proposent de faciliter l'apprentissage du réseau de neurones convolutif en représentant les différents points anatomiques du corps humain en volumes appelés voxels plutôt qu'en coordonnées ponctuelles x, y, z bien précises. Moon *et al*. (2020) utilisent une approche similaire à Pavlakos *et al*. (2017), mais estiment plutôt la pose et la forme.

1.3.2 Approche par optimisation (ou multi-étapes)

L'approche par optimisation consiste à effectuer dans un premier temps une estimation de pose 2D, puis à déformer un modèle paramétrique tridimensionnel de manière à ce que sa projection dans le plan de l'image corresponde aux points-clés 2D. Les coordonnées 3D sont ainsi obtenues (Figure 1.21). Les méthodes multi-étapes sont sensibles aux erreurs de l'estimation de pose 2D. De plus, elles ne tiennent pas compte d'informations potentiellement utiles pouvant être tirées de l'image elle-même comme les contours ou les contrastes. Toutefois, étant donné que les estimateurs de pose 2D sont de plus en plus

robustes, à l'heure de la rédaction de ce mémoire, l'approche par optimisation semble plus performante que l'approche par régression puisque cette dernière constitue un problème encore trop complexe d'un point de vue computationnel.

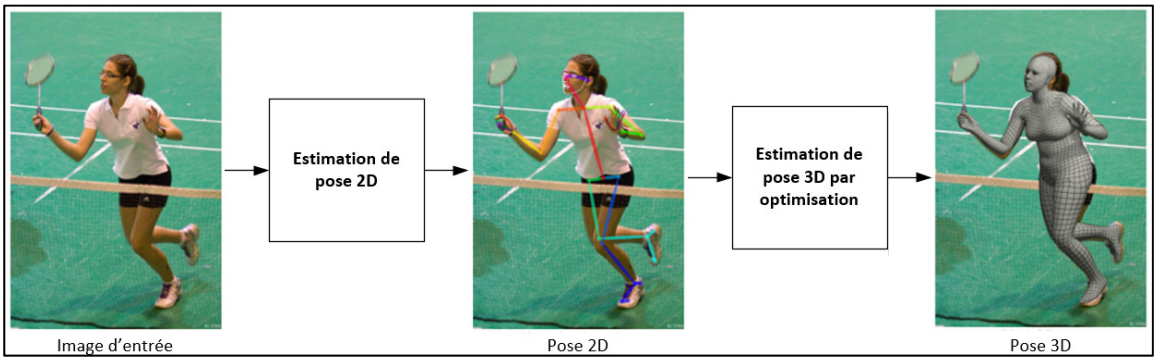


Figure 1.21 Processus d'estimation de pose 3D par optimisation
Adaptée de Godwisdom (2020)

Suivant l'approche par optimisation, Bogo *et al*. (2016) utilisent un CNN (Pischulin *et al*., 2016) pour faire d'abord l'estimation de pose 2D, puis minimisent une fonction objectif pour déterminer la pose 3D en tenant compte de l'erreur de reprojection du modèle sur l'image et de la plausibilité de la pose. Cette dernière peut être évaluée entre autres grâce à l'utilisation d'un modèle pose + forme, puisque celui-ci permet le calcul de l'interpénétration du corps. Cheng *et al*. (2020) utilisent l'estimateur de pose 2D HRNet (Sun *et al*., 2019), aussi un CNN, pour obtenir la pose bidimensionnelle. Puis, ils optent pour un réseau convolutif temporel (Lea *et al*., 2017) afin d'apprendre les dépendances temporelles entre des poses consécutives. Un discriminateur cinématique spatio-temporel pénalise les changements trop abrupts de la longueur des membres ou de l'angle entre ceux-ci, et des masques et l'ajout de bruit lors de l'entraînement de leur modèle le rendent plus robuste aux occlusions. Tout comme les autres méthodes d'approche par optimisation présentées jusqu'ici, Hu *et al*. (2021) utilisent un CNN pour estimer la pose 2D. Ils expérimentent ainsi avec HRNet et CPN (Chen *et al*., 2017). Les poses sont par la suite mises sous forme de séquences de graphes. Celles-ci sont données à leur modèle U-CondDGConv, qui est un réseau convolutif de

graphes conditionnel entraîné à les estimer dans l'espace 3D. Zhang *et al*. (2022) utilisent plutôt le réseau transformeur introduit par Vaswani *et al*. (2017) pour cette tâche.

1.3.3 Approche mixte

Kolotouros *et al*. (2019) proposent SPIN, un modèle basé sur un CNN et qui combine l'approche par régression et l'approche par optimisation dans le but de bénéficier de leurs avantages respectifs. Puisque cette dernière méthode ne s'applique qu'aux images, Kocabas *et al*. (2020) l'adaptent pour créer VIBE, un estimateur de pose 3D capable de traiter des vidéos sur lesquelles peuvent figurer plusieurs personnes. À travers PARE, Kocabas *et al*. (2021) s'attaquent au problème d'occlusion en entraînant le modèle à prédire des poses où certaines parties du corps sont masquées.

Le tableau 1.6 compare les différentes méthodes d'estimation de pose 3D présentées dans cette section. Parmi celles-ci, peu testent leur modèle sur l'ensemble de données HumanEva. Ceci s'explique par le fait qu'il est le plus ancien. Comme mentionné à la section 1.3.1, d'autres ensembles de données plus larges ont depuis été constitués, et la plupart des auteurs préfèrent présenter leurs résultats sur ceux-ci. À l'inverse, les résultats sur 3DPW sont également plus rares, puisque sa publication est la plus récente. Les auteurs d'articles datant d'avant 2018 ne pouvaient donc pas s'en servir à des fins d'évaluation. 3DPW possédant aussi la vérité terrain de forme, ce sont généralement les travaux concernant les estimateurs de pose + forme qui publient des résultats sur cet ensemble de données. Ainsi, Human3.6M et MPI-INF-3DHP sont les ensembles de données pour lesquels la majorité des méthodes présentées dans cette section sont évaluées. En observant le tableau 1.6, il est possible de voir que les estimateurs de pose 3D suivant l'approche par optimisation sont plus performants, et ce, autant dans le cas de Human3.6M que de MPI-INF-3DHP. L'approche mixte, qui possède des caractéristiques de l'approche multi-étapes, présente des performances inférieures à cette dernière, mais généralement plus élevées que l'approche par régression.

Tableau 1.6 Comparaison de différents estimateurs de pose 3D

Auteurs	Type d'approche	Type de modèle	Performances sur différents ensembles de données							
			HumanEva	Human3.6M		MPI-INF-3DHP			3DPW	
			MPJPE (↓) (mm)	MPJPE (↓) (mm)		PCK (↑) (%)	AUC (↑) (%)	MPJPE (↓) (mm)	MPJPE (↓) (mm)	
				P1	P2				P1	P2
Zhou *et al*. (2016)	Régression	CNN	-	-	107,26	-	-	-	-	-
Pavlakos *et al*. (2017)	Régression	CNN	24,3	71,90	51,9	-	-	-	-	-
Moon *et al*. (2020)	Régression	CNN	-	55,7	41,1	-	-	-	93,2	57,7
Bogo *et al*. (2016)	Optimisation	CNN	-	82,3	-	-	-	-	-	-
Cheng *et al*. (2020)	Optimisation	TCN	**13,5**	40,1	30,7	84,1	-	-	-	-
Hu *et al*. (2021)	Optimisation	GCN	-	41,1	32,1	**97,9**	**69,5**	**42,5**	-	-
Zhang *et al*. (2022)	Optimisation	Transformeur	20,9	**39,8**	**30,6**	94,4	66,5	54,9	-	-
Kolotouros *et al* (2019)	Mixte	CNN	-	62,0	41,1	76,4	37,1	105,2	-	-
Kocabas *et al* (2020)	Mixte	CNN	-	65,9	41,5	89,0	-	97,7	82,9	51,9
Kocabas *et al*. (2021)	Mixte	CNN	-	-	-	-	-	-	**74,5**	**46,5**

Légende :

MPJPE : *Mean Per Joint Position Error* PCK : *Percentage of Correct Keypoints* AUC : *Area Under Curve*

1.4 Reconnaissance d'actions

La reconnaissance d'actions est également une tâche à laquelle peut être consacré un réseau de neurones. Il s'agit pour ce dernier d'identifier les actions qu'effectuent une ou plusieurs personnes sur, généralement, une séquence vidéo.

1.4.1 Reconnaissance d'actions sur des images

Plusieurs ensembles de données ont été élaborés, dont les reconnus UCF101 (Soomro *et al*., 2012) et Kinetics-600 (Carreira *et al*., 2018), dans le but d'entraîner des modèles à reconnaître différentes activités humaines à partir d'images. UCF101 est l'extension de UCF50 (Reddy & Shaw, 2013). Il couvre 101 actions pouvant être regroupées sous les cinq thèmes suivants : interaction humain-objet, mouvement du corps, interaction humain-humain, jouer d'un instrument de musique et sports. Cet ensemble de données contient 13 320 extraits issus de 2 500 vidéos provenant de la plateforme YouTube, d'une longueur moyenne de 7,21 secondes (durée la plus courte de 1,06 seconde et durée la plus longue de 71,04 secondes). Les actions d'intérêt en boxe sont : coup de poing, boxe sur sac de frappe et boxe sur sac de frappe de vitesse. La liste exhaustive des actions se trouve en ANNEXE II. Kinetics-600 est l'extension de Kinetics-400 (Kay *et* al., 2017) regroupant 600 classes d'actions. Kinetics-600 est constitué de 495 547 extraits issus d'autant de vidéos provenant de la plateforme YouTube. En plus d'avoir davantage d'actions que UCF101, cet ensemble de données possède plus de variabilité puisque chaque extrait provient d'une vidéo différente, tandis que UCF101 peut contenir des extraits d'une personne effectuant plusieurs fois la même action. Étant donné le très grand nombre d'actions de cet ensemble de données, la liste complète n'est pas fournie dans ce mémoire. Le lecteur est cependant invité à consulter les publications de Kay *et al*. (2017) et Carreira *et al*. (2018) pour connaître les différentes actions y figurant. Les thèmes principaux couverts par l'ensemble de données Kinetics sont :

1) Artisanat
2) Sauts en athlétisme
3) Tirs et lancers en athlétisme
20) Sports en hauteur
21) Interactions avec des animaux
22) Jongler

4) Maintenance automobile
5) Sports de balles et de ballons
6) Mouvements du corps
7) Tâches ménagères
8) Interaction avec des vêtements
9) Communication
10) Cuisiner
11) Danser
12) Boire et manger
13) Utiliser du matériel électronique
14) Jardinage
15) Golf
16) Gymnastique
17) Interaction avec des cheveux
18) Actions avec les mains
19) Mouvements de la tête et de la bouche
23) Maquillage
24) Arts martiaux
25) Actions diverses
26) Déplacements terrestres
27) Déplacements aquatiques
28) Musique
29) Actions avec du papier
30) Hygiène personnelle
31) Jeux
32) Sports de raquette ou de batte
33) Actions sur la neige ou sur la glace
34) Nager
35) Contact avec quelqu'un
36) Outils
37) Sports aquatiques
38) Épilation

Des méthodes basées sur des réseaux de neurones convolutifs bâtis de zéro (Diba *et al*., 2020) ou préétablis comme le ResNet de He *et al*. (2016) (Qiu *et al*., 2019; Li *et al*., 2021) obtiennent des résultats de classification presque parfaits sur l'ensemble de données UCF101 (Tableau 1.7). Les performances sur Kinetics-600 sont moins saturées, avec une exactitude de classification avoisinant les 80 %. Étant donné que cet ensemble de données contient environ six fois plus d'actions que le premier, cela rend la tâche de classification plus ardue, ce qui explique les performances moins élevées des modèles de l'état des connaissances sur Kinetics-600.

Tableau 1.7 Performances des classificateurs sur UCF101 et Kinetics-600

Auteurs	Nom du modèle	Exactitude (%)	
		UCF101	Kinetics-600
Qiu *et al*., 2019	LGD-3D	**98,2**	**83,1**
Diba *et al*., 2020	HATNet	97,8	81,6
Li *et al*., 2021	PERF-Net	**98,2**	82,0

1.4.2 Reconnaissance d'actions à partir des poses

Utiliser les poses humaines pour la reconnaissance d'actions plutôt que des images permet d'obtenir un modèle invariant aux changements de luminosité ou de contraste sur les images, ainsi qu'aux différents points de vue de la caméra. Les séquences de poses 3D détiennent également une plus grande richesse d'information, puisqu'elles permettent d'obtenir la cinématique de différentes actions. En utilisant les poses consécutives d'un même mouvement, il est ainsi possible de calculer la vitesse et l'accélération des différentes parties du corps, ce qui peut aider à faire la distinction entre certains types d'actions. De plus, les poses 3D permettent d'obtenir l'information de profondeur, qui n'est pratiquement pas accessible avec des images uniquement.

À l'heure de la rédaction de ce mémoire, le plus large ensemble de données pour l'entraînement et l'évaluation des modèles de reconnaissance d'actions à partir des poses est NTU RGB+D (Shahroudy *et al*., 2016). Celui-ci contient 60 actions différentes ayant comme trois thèmes principaux les activités quotidiennes (boire de l'eau, saluer, etc.), les conditions médicales (éternuer, tomber, etc.) et les interactions entre deux personnes (coup de poing, coup de pied, etc.). Pour enregistrer les poses 3D, les auteurs utilisent le capteur de mouvement Microsoft Kinect v2. L'ensemble de données inclut des extraits vidéos d'une durée moyenne d'environ 2 secondes pour chaque action, en plus des cartes de profondeur, des poses 3D et des images infrarouges associées. En 2020, Liu *et al*. prolongent cet ensemble de données en y ajoutant 60 actions supplémentaires pour former NTU RGB+D 120. Ils utilisent le même type de capteur de mouvement. Les listes complètes des actions

figurant dans les ensembles de données NTU RGB+D et NTU RGB+D 120 sont présentées à l'ANNEXE II. Bien qu'une action typique de la boxe – coup de poing – y figure, il est à noter que cela n'est pas suffisant pour entraîner un modèle à reconnaître les différentes techniques offensives de ce sport.

Jusqu'en 2018, les réseaux de neurones les plus performants sur NTU RGB+D sont majoritairement de type longue mémoire à court terme. Shahroudy *et al*. (2016), instigateurs de l'ensemble de données NTU RGB+D, utilisent un LSTM pour la classification d'actions par les poses. Ils proposent de représenter le corps humain en cinq parties, soient le torse, les deux bras et les deux jambes, qui sont chacune traitée individuellement par une unité LSTM liée à une porte de sortie commune. Zhang *et al*. (2017) entraînent des sous-réseaux LSTM à identifier le meilleur point d'observation de la pose pour optimiser la reconnaissance d'actions. Les auteurs utilisent ensuite un réseau LSTM dédié à la reconnaissance d'actions pour traiter les coordonnées 3D finalement obtenues.

Puis, en 2018, les réseaux convolutifs de graphes font leur apparition dans le domaine de la reconnaissance d'actions à partir des poses pour finalement dominer l'état des connaissances. Il est particulièrement pertinent d'utiliser les graphes dans un contexte de reconnaissance d'actions par les poses, puisque le corps humain peut lui-même être représenté par un graphe, où les différentes articulations en sont les nœuds, et où les membres en sont les segments (Figure 1.22). Yan *et al*. (2018) sont les premiers à utiliser un GCN pour la reconnaissance d'actions par les poses. Ils modélisent un graphe spatial lié à la structure du corps humain, ainsi qu'un graphe temporel pour l'interaction des parties du corps lors des mouvements. Cheng *et al*. (2020) introduisent une opération de décalage du graphe. Celle-ci permet l'échange de caractéristiques entre tous les nœuds du graphe pour qu'à chaque étape, le réseau détienne toute l'information du corps humain.

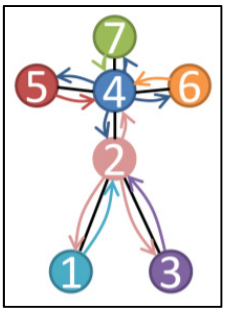


Figure 1.22 Graphe modélisant la pose humaine
Tirée de Cheng *et al*. (2020)

Quelques tentatives ont également été réalisées dans le but de combiner les avantages des GCN avec ceux des LSTM. Si *et al*. (2019) exploitent à la fois un GCN pour traiter l'information spatiale, et un LSTM pour tirer profit de l'information temporelle. De manière similaire, Zhao *et al*. (2019) se servent d'un GCN combiné à un LSTM pour traiter ensemble l'information spatiale et l'information temporelle, et considèrent les paramètres de ces modèles comme des variables aléatoires dans un problème statistique d'inférence bayésienne afin de prédire l'action survenant sur une séquence de poses. Bien que les méthodes combinant les GCN aux LSTM cherchent à bénéficier des avantages respectifs de ces deux types de réseaux de neurones, il est à noter qu'au moment de la rédaction de ce mémoire, les réseaux convolutifs de graphes présentent généralement de meilleures performances de classification que les LSTM seuls et que les combinaisons GCN-LSTM. Le tableau 1.8 rassemble les performances des différents modèles évoqués dans cette section sur l'ensemble de données NTU RGB+D. Il est à noter que les performances inter-vue sont supérieures à celles inter-sujet. Dans le premier cas, le même sujet se retrouve à la fois dans les données d'entraînement et d'évaluation, ce qui permet une meilleure classification d'actions. Il semble donc que la variabilité inter-sujet pour un même mouvement soit suffisamment forte pour affecter les performances du modèle.

Tableau 1.8 Performances des classificateurs sur l'ensemble NTU RGB+D

Auteurs	Méthode	Type de modèle	Exactitude (%)	
			Inter-vue	Inter-sujet
Shahroudy *et al*. (2016)	P-LSTM	LSTM	70,3	62,9
Zhang *et al*. (2017)	VA-LSTM	LSTM	87,6	79,4
Yan *et al*. (2018)	ST-GCN	GCN	88,3	81,5
Cheng *et al*. (2020)	Shift-GCN	GCN	**96,5**	**90,7**
Si *et al*. (2019)	AGC-LSTM	GCN-LSTM	95,0	89,2
Zhao *et al*. (2019)	Bayesian GC-LSTM	GCN-LSTM	89,0	81,8

La présente revue de la littérature a permis de mettre en lumière le besoin d'extraire automatiquement des métriques de performance en boxe telles que le nombre de coups donnés dans un combat ainsi que le contrôle du ring par les duellistes. Bien que des méthodes antérieures aient réussi à classifier par apprentissage machine certaines techniques offensives effectuées lors de séances de *shadow boxing*, elles nécessitent de l'équipement spécialisé qui rend leur application difficile pour l'analyse de réels combats. Les avancées en vision par ordinateur permettent d'extraire les poses 3D à partir de vidéos uniquement, ce qui rend ce type de données d'entrée attrayant pour l'entraînement d'un classificateur. Ces idées constituent ainsi la ligne directrice de la méthodologie suivie dans le cadre de ce projet. Celle-ci est détaillée dans le chapitre 3.

CHAPITRE 2

PROBLÉMATIQUE, OBJECTIFS ET RETOMBÉES

2.1 Problématique

Des études ont démontré que le nombre de coups donnés dans un combat de boxe et leur efficience sont des métriques importantes dans l'analyse des performances d'un athlète et ses chances d'obtenir un verdict de victoire (Ashker *et al*., 2011; Davis *et al*., 2015; Thompson & Lamb, 2016). De fait, ces éléments font partie intégrante du système de pointage adopté par l'Association Internationale de Boxe. Celui-ci inclut également des composantes stratégiques, dont notamment le contrôle du combat par chaque athlète. Ceci peut se traduire par l'occupation du ring, par exemple, si l'un des boxeurs parvient à coincer son adversaire entre lui et les cordes afin de limiter ses mouvements. Il s'agit donc de statistiques pertinentes à détenir pour les équipes, autant pour évaluer et mieux entraîner leurs propres boxeurs que pour analyser l'opposition. À l'heure de la rédaction de ce mémoire, pour recueillir ces métriques, les équipes d'excellence en boxe affiliées à l'Institut national du sport du Québec doivent visionner et annoter manuellement les vidéos de combat. Il s'agit d'un processus requérant une certaine expertise et demandant beaucoup de temps.

Certaines tentatives d'automatisation de reconnaissance et de comptabilisation de techniques offensives en boxe ont été faites. Toutefois, elles nécessitent du matériel spécialisé comme un système de capture de mouvement Vicon (Khoury & Liu, 2009; Soekarjo *et al*., 2019), une caméra à temps de vol (Kasiri-Bidhendi *et al*., 2015) ou des capteurs inertiels (Soekarjo *et al*., 2019; Worsey *et al*., 2020; Khasanshin, 2021).

Il est possible de tirer profit des récentes avancées en estimation de pose 3D pour contourner ce besoin matériel. En effet, ces réseaux de neurones peuvent produire des résultats près de ceux d'un système de capture de mouvement traditionnel, en ne requérant en entrée que des

vidéos enregistrées par une simple caméra. Cela permet du même coup de développer une méthode simple à déployer lors de combats à l'étranger, ou pouvant même être utilisée sur des combats passés dont les vidéos sont déjà enregistrées.

2.2 Objectifs du projet

L'objectif principal de cette maîtrise est de développer une méthode d'extraction automatique de métriques de performance en boxe basée sur la vision par ordinateur capable d'identifier et de comptabiliser deux positions et six principales techniques offensives en boxe sur une vidéo de *shadow boxing*, soit la garde ouverte, la garde, le direct avant, le direct arrière, le crochet avant, le crochet arrière, l'uppercut avant et l'uppercut arrière. L'évaluation quantitative de cette méthode fait également l'objet de ce projet. Cette dernière doit aussi faire le suivi des boxeurs sur l'aire de combat afin d'éventuellement analyser son occupation par les deux athlètes.

L'Institut national du sport du Québec souhaite éventuellement posséder un outil d'analyse automatique des performances ayant la capacité de reconnaître adéquatement 90 % des coups lancés par un boxeur. Cette dernière valeur est considérée comme une cible idéale à atteindre, mais 80 % représentent également une exactitude de classification que les experts de l'Institut estiment satisfaisante.

Afin de remplir l'objectif principal, les objectifs secondaires suivants doivent être atteints :

1) Constituer un ensemble de données spécifiques à la boxe permettant l'entraînement d'un classificateur pour la reconnaissance des huit actions suivantes : garde ouverte, garde, direct avant, direct arrière, crochet avant, crochet arrière, uppercut avant et uppercut arrière;
2) Déterminer les paramètres du classificateur permettant d'obtenir les meilleures performances de classification sur l'ensemble de données ainsi constitué;
3) Développer un algorithme de suivi des boxeurs sur l'aire de combat grâce à la vision par ordinateur.

2.3 Délimitations du projet

La portion de la méthode concernant la reconnaissance d'actions se limite à effectuer cette dernière sur des vidéos de *shadow boxing*. Ceci permet d'éviter l'occlusion entre deux personnes, et ainsi de s'assurer d'avoir des données d'entrée idéales où l'entièreté de l'athlète est visible. La contrainte d'un point de vue fixe est imposée, autant pour la reconnaissance d'actions et le suivi des boxeurs sur l'aire de combat. De plus, la reconnaissance d'actions en temps réel, ainsi que le suivi des boxeurs en temps réel ne font pas partie de ce projet de maîtrise.

2.4 Retombées attendues et importance de l'étude

À la connaissance de l'auteure de ce mémoire, la présente méthode d'extraction automatique de métriques de performance en boxe est la première à n'utiliser que des vidéos en entrée pour la reconnaissance de techniques spécifiques à ce sport grâce à l'apprentissage machine. D'autres tentatives d'automatisation antérieures nécessitaient plutôt l'enregistrement de données provenant d'équipement spécialisé tel qu'un système de capture de mouvement, une caméra à temps de vol ou des capteurs inertiels.

De plus, la méthode développée dans le cadre de cette maîtrise est un premier pas vers l'analyse automatique de combats de boxe, ce qui est particulièrement d'attrait pour les entraîneurs des équipes affiliées à l'Institut national du sport du Québec. À l'heure actuelle, ceux-ci doivent visionner et annoter manuellement les différents événements survenant durant un combat, ce qui est un processus très long et devant être effectué par un expert analyste des performances. La méthode d'extraction automatique de métriques de performance en boxe offrira donc aux entraîneurs un moyen de détenir rapidement et automatiquement des informations utiles à l'évaluation des boxeurs, telles que le nombre de coups donnés par chacun d'eux, ainsi que leur efficience. Elle permettra également de quantifier l'occupation de l'aire de combat par les athlètes.

CHAPITRE 3

MÉTHODOLOGIE

Le développement de la méthode automatisée d'extraction de métriques de performance en boxe nécessite plusieurs étapes, qui sont illustrées sur la figure 3.1 et détaillées dans ce chapitre.

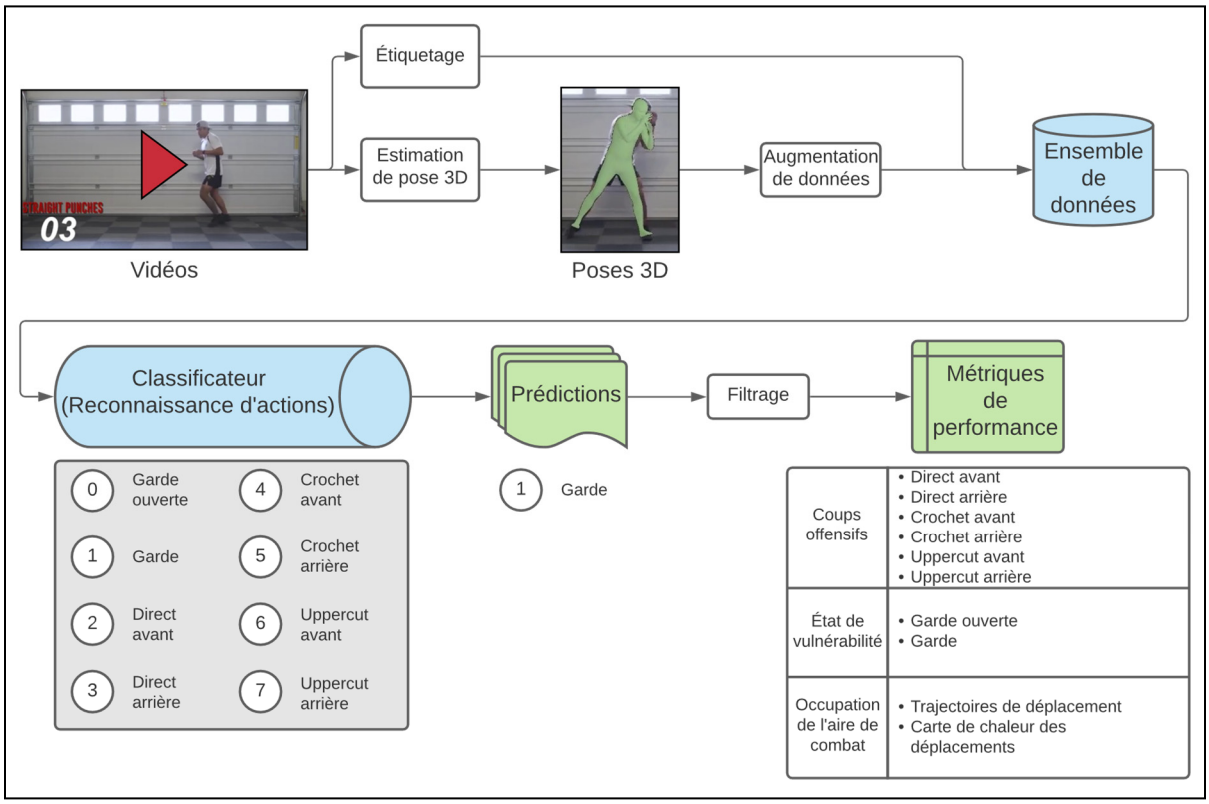


Figure 3.1 Schéma de la méthode

3.1 Base de données pour l'entraînement du classificateur

En consultant la littérature, il est possible de constater qu'il n'existe aucune base de données spécifique à la boxe et dédiée à l'entraînement de classificateurs à l'aide de poses. Tout comme les auteurs des méthodes présentées à la section 1.1.3 de ce mémoire, il faut ainsi constituer une base de données de poses 3D annotées manuellement afin d'entraîner le

classificateur à reconnaître les huit positions et techniques offensives d'intérêt dans ce projet. Les sous-étapes pour arriver à cette fin sont décrites dans cette section.

3.1.1 Sélection des vidéos

Dans le cadre de cette maîtrise, les vidéos d'intérêt pour développer la méthode sont celles ne contenant qu'un seul boxeur, soit du *shadow boxing*. Ce choix est fait pour éviter le phénomène d'occlusion entre plusieurs personnes (Figure 3.2) afin d'assurer la qualité des poses extraites. Dans le même but, aucune vidéo où l'athlète fait dos à la caméra n'est retenue. Il peut cependant être de face ou de côté. L'occlusion, autant entre deux personnes que par les parties du corps d'un même individu, est un défi actuel dans le domaine de l'estimation de pose 3D et engendre des erreurs dans les résultats que les modèles produisent. Divers travaux tentent de rendre les estimateurs de pose 3D robustes à l'occlusion (Cheng *et al.*, 2019; Cheng *et al.*, 2020; Kocabas *et al.*, 2021), mais cette problématique demeure présente à l'heure actuelle de la rédaction de ce mémoire. L'amélioration de l'estimation de pose 3D ne faisant pas partie de la portée du présent projet, les vidéos sont choisies afin d'obtenir les meilleurs résultats possible des estimateurs de pose 3D déjà existants.

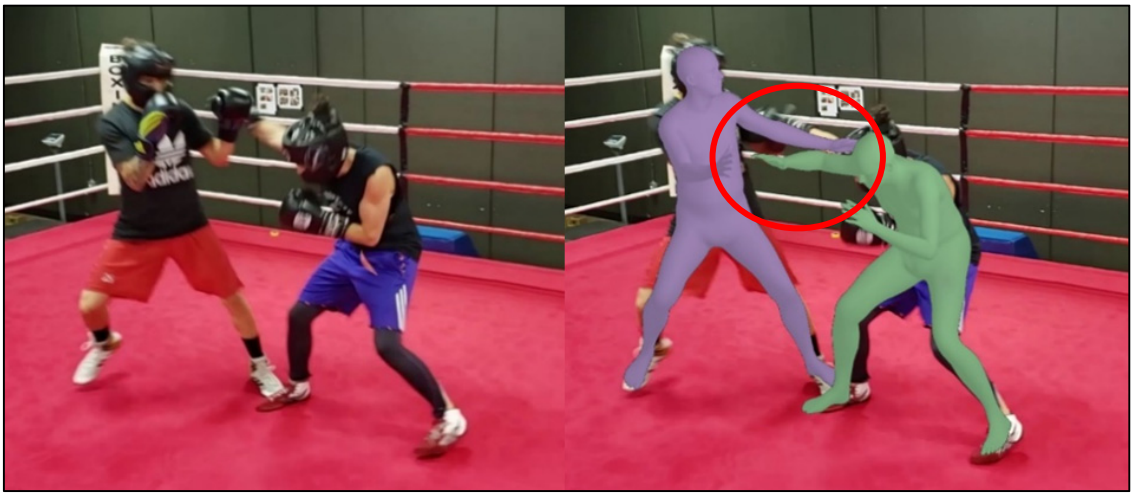


Figure 3.2 Erreur d'estimation de pose 3D liée à l'occlusion entre deux personnes

Les extraits vidéos peuvent contenir différents types de positions et techniques offensives, et plusieurs exemples de celles-ci.

Deux sources permettent d'obtenir de telles vidéos, soit la plateforme YouTube et l'Institut national du sport du Québec. L'ANNEXE III présente une liste des vidéos YouTube utilisées dans l'ensemble de données constitué dans le cadre de ce présent projet. Par l'entremise d'un certificat éthique, l'INS Québec a pu transférer pour cette maîtrise les vidéos de *shadow boxing* d'athlètes de haut niveau. Le tableau 3.1 résume les caractéristiques des vidéos retenues pour constituer la base de données.

Tableau 3.1 Résumé des données vidéo

Techniques de boxe	8	234 directs avant 192 directs arrière 249 crochets avant 148 crochets arrière 127 uppercuts avant 146 uppercuts arrière 18 gardes ouvertes 1 065 gardes	(837 images) (695 images) (1 345 images) (684 images) (554 images) (600 images) (1 045 images) (18 942 images)
Personnes	22	15 YouTube 7 INS Québec	
Extraits	187	154 YouTube 33 INS Québec	
Longueur moyenne d'un extrait	5,31 secondes	min : 0,53 seconde max : 30,53 secondes	
Durée totale	16,5 minutes		
Fréquence d'images	30 images/seconde		
Résolution	1920 x 1080 pixels		

Dans le but d'éventuellement effectuer une validation croisée à 5 plis, les données sont séparées de façon à obtenir 5 ensembles d'évaluation différents. La répartition se fait sous les contraintes suivantes :

1) L’ensemble d’évaluation contient au moins 100 exemples de chacun des six types de coups de poing;
2) Les ensembles d’entraînement, de validation et d’évaluation résultants contiennent respectivement l’entièreté des vidéos d’une même personne, afin d’assurer qu’un athlète ne se retrouve pas à la fois dans les données d’entraînement et de validation/évaluation.

Le nombre d’images et d’exemples dans chaque ensemble est donné dans le tableau 3.2. Les valeurs moyennes résultantes sont indiquées dans le tableau 3.3.

Tableau 3.2 Nombre d'exemples (images) pour la validation croisée à 5 plis

Classe	1			2			3		
	Entraînement	Validation	Évaluation	Entraînement	Validation	Évaluation	Entraînement	Validation	Évaluation
Garde ouverte	21 (1 185)	1 (54)	2 (27)	21 (1 079)	1 (54)	2 (133)	21 (1 207)	1 (54)	2 (5)
Garde	928 (15 637)	124 (2 023)	124 (2 914)	772 (16 240)	124 (2 023)	280 (2 311)	942 (14 657)	124 (2 023)	110 (3 894)
Direct avant	208 (745)	31 (98)	17 (50)	170 (609)	31 (98)	55 (186)	205 (734)	31 (98)	20 (61)
Direct arrière	192 (694)	14 (54)	11 (48)	130 (524)	14 (54)	73 (218)	191 (672)	14 (54)	12 (70)
Crochet avant	239 (1 275)	18 (83)	11 (76)	148 (822)	18 (83)	102 (529)	221 (1 131)	18 (83)	29 (220)
Crochet arrière	114 (501)	17 (79)	17 (104)	113 (544)	17 (79)	18 (61)	112 (508)	17 (79)	19 (97)
Uppercut avant	107 (471)	9 (60)	24 (99)	71 (298)	9 (60)	60 (272)	121 (533)	9 (60)	10 (37)
Uppercut arrière	114 (461)	16 (60)	29 (129)	87 (342)	16 (60)	56 (248)	132 (547)	16 (60)	11 (43)

Classe	4			5		
	Entraînement	Validation	Évaluation	Entraînement	Validation	Évaluation
Garde ouverte	19 (1 124)	1 (54)	4 (88)	16 (1 054)	1 (54)	7 (158)
Garde	805 (15 049)	124 (2 023)	247 (3 502)	867 (14 717)	124 (2 023)	185 (3 834)
Direct avant	167 (617)	31 (98)	58 (178)	176 (575)	31 (98)	49 (220)
Direct arrière	129 (485)	14 (54)	74 (257)	183 (666)	14 (54)	20 (76)
Crochet avant	205 (1 146)	18 (83)	45 (205)	195 (1 091)	18 (83)	55 (260)
Crochet arrière	98 (471)	17 (79)	33 (134)	121 (561)	17 (79)	10 (44)
Uppercut avant	120 (517)	9 (60)	11 (53)	121 (528)	9 (60)	10 (42)
Uppercut arrière	132 (548)	16 (60)	11 (42)	107 (462)	16 (60)	36 (128)

Tableau 3.3 Nombre d'exemples (images) moyen pour la validation croisée à 5 plis

Classe	**Moyenne**		
	Entraînement	Validation	Évaluation
Garde ouverte	20 (1 130)	1 (54)	3 (82)
Garde	863 (15 260)	124 (2 023)	189 3 291
Direct avant	185 (656)	31 (98)	40 (139)
Direct arrière	165 (608)	14 (54)	38 (134)
Crochet avant	202 (1 093)	18 (83)	48 (258)
Crochet arrière	112 (517)	17 (79)	19 (88)
Uppercut avant	108 (469)	9 (60)	23 (101)
Uppercut arrière	114 (472)	16 (60)	29 (118)

3.1.2 Étiquetage des images

L'étiquetage consiste à visionner les vidéos décrites à la section ci-dessus image par image, et à attribuer à chacune d'elle une étiquette qui correspond à l'action qui y est représentée. Ce processus est effectué par l'auteure de ce mémoire. Les huit classes d'actions possibles sont la garde ouverte, la garde, le direct avant, le direct arrière, le crochet avant, le crochet arrière, l'uppercut avant et l'uppercut arrière. L'ANNEXE IV présente des exemples visuels de la façon dont sont étiquetées les images pour les différentes classes d'action.

3.1.3 Estimation de pose 3D

Les vidéos décrites à la section 3.1.1 sont ensuite données en entrée à un estimateur de pose 3D. Pour choisir le modèle à utiliser parmi ceux de la littérature, deux critères sont d'abord établis :

1) L'estimateur de pose 3D doit faire partie de l'état des connaissances;
2) Son code doit être public, et il doit permettre l'entraînement et l'utilisation sur de nouvelles données.

D'une part, cela permet de s'assurer que l'estimateur de pose 3D sélectionné est performant. D'autre part, il est nécessaire que le code soit public afin de pouvoir utiliser le modèle et le réentraîner au besoin avec des données spécifiques à la boxe. En 2020, année du début de ce projet de maîtrise, trois modèles potentiels satisfaisant les critères ci-dessus ont été sélectionnés, soient MargiPose, VideoPose3D et SPIN. Leurs performances sur les ensembles de données Human3.6M et MPI-INF-3DHP sont présentées dans le tableau 3.4.

Tableau 3.4 Comparaison de MargiPose, VideoPose3D et SPIN

Ensemble de données			**Estimateur de pose 3D**		
			MargiPose (Nibali *et al*., 2019)	**VideoPose3D** (Pavllo *et al*., 2019)	**SPIN** (Kolotouros *et al*., 2019)
			Régression	Optimisation	Mixte
Human3.6M	**MPJPE (↓) (mm)**	P1	55,4	**46,8**	62,0
		P2	39,0	**36,5**	41,1
MPI-INF-3DHP	**PCK (↑) (%)**		**85,4**	-	76,4
	AUC (↑) (%)		**91,3**	-	37,1
	MPJPE (↓) (mm)		**47,0**	-	105,2

Légende :

MPJPE : *Mean Per Joint Position Error* PCK : *Percentage of Correct Keypoints*

AUC : *Area Under Curve*

Bien que cela ne soit pas énoncé explicitement dans les articles d'origine, des expérimentations réalisées dans le cadre de cette maîtrise ont permis de constater que sur des images spécifiques au sport, les méthodes utilisant le modèle pose + forme SMPL (comme SPIN, VIBE et PARE) génèrent de manière constante des poses plausibles et stables, tandis que d'autres méthodes de l'état des connaissances souffrent de l'instabilité des poses. Ce phénomène est bien connu dans le domaine de l'estimation de pose 3D, et décrit le cas de figure où la pose semble adéquate dans le plan de l'image d'entrée, tandis qu'une fois tournée autour de l'axe longitudinal, elle se révèle physiquement impossible à maintenir. Par exemple, la figure 3.3 compare les trois estimateurs de pose 3D sélectionnés sur une image d'un boxeur. Tous produisent des résultats satisfaisants lorsqu'on les observe dans le plan de l'image. Toutefois, il est possible de voir que lorsqu'une rotation est effectuée, la pose donnée par VideoPose3D n'est pas physiquement possible en raison de son inclinaison. Le résultat donné par SPIN est quant à lui plausible, peu importe l'angle d'observation.

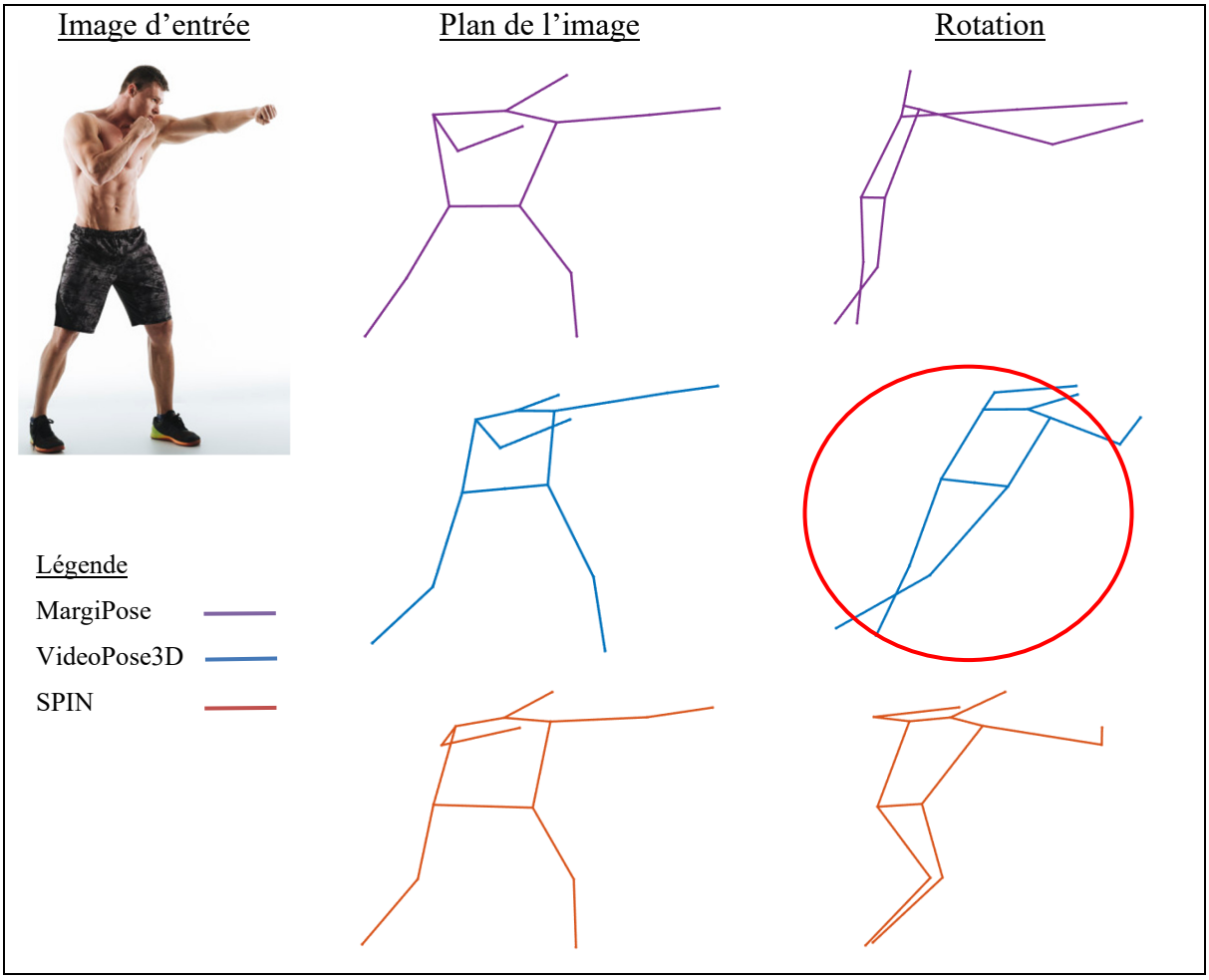


Figure 3.3 MargiPose, VideoPose3D et SPIN sur une image d'un boxeur

Sur une photo d'un joueur de soccer, c'est plutôt MargiPose qui produit une pose instable et physiquement incorrecte au niveau du genou droit lorsque regardée d'un point de vue différent de celui de l'image d'entrée (Figure 3.4).

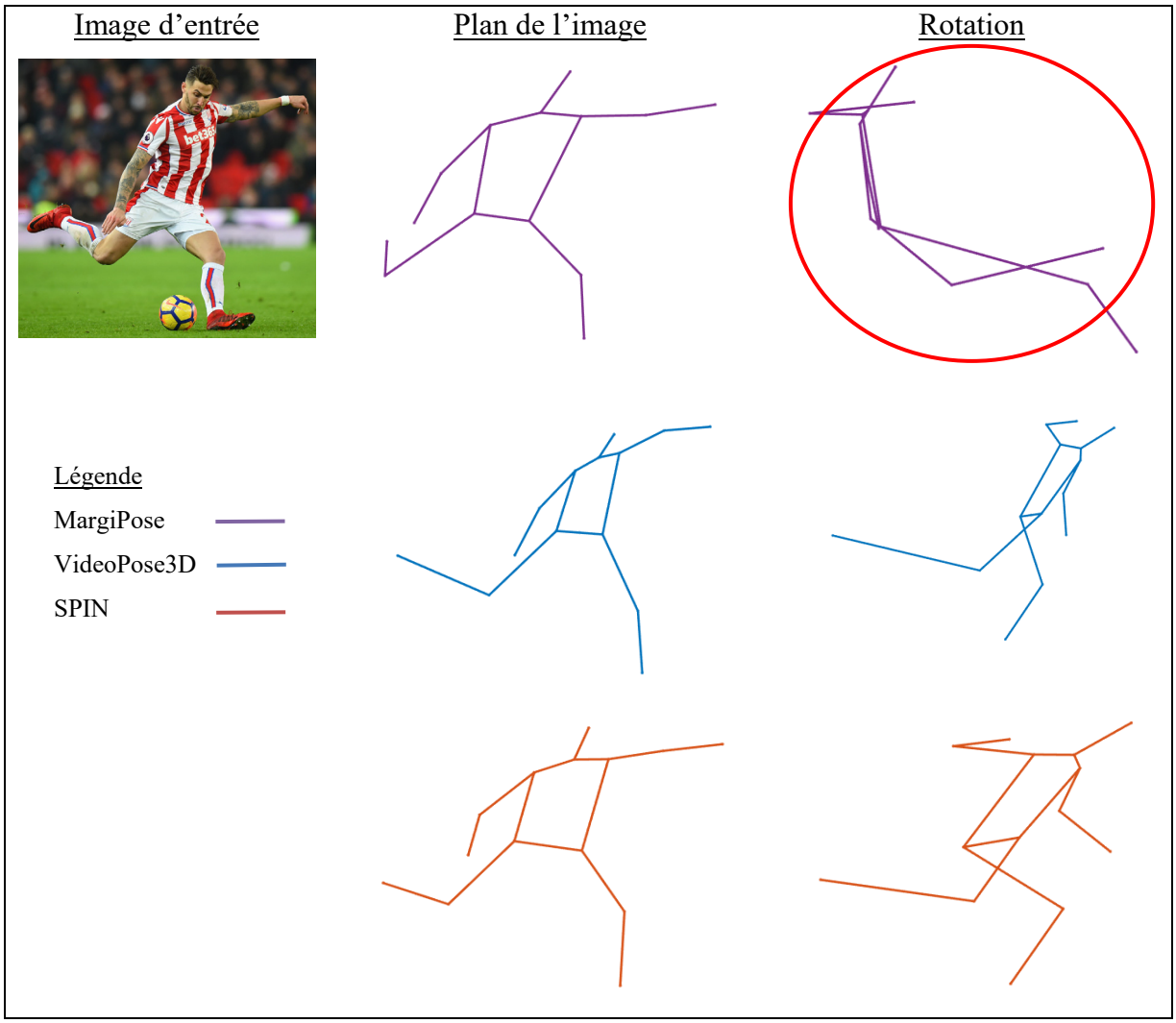


Figure 3.4 MargiPose, VideoPose3D et SPIN sur une image d'un joueur de soccer

Comme ils sont l'adaptation de SPIN pour des vidéos, VIBE et PARE sont les estimateurs de pose 3D sélectionnés pour générer les poses servant à constituer l'ensemble de données spécifique à la boxe. Ces deux modèles sont développés par la même équipe de recherche (Institut Max-Planck pour les systèmes intelligents). VIBE est utilisé pour traiter les vidéos provenant de la plateforme YouTube, tandis que PARE est adopté pour les vidéos rendues disponibles par l'INS Québec. Ce dernier modèle ayant été spécifiquement entraîné pour être robuste à l'occlusion, il permet d'obtenir de meilleurs résultats sur les vidéos de l'aire de combat située à l'Institut national du sport du Québec, dont les cordes peuvent provoquer à

certains moments des erreurs liées à l'occlusion. Par exemple, la figure 3.5 présente une erreur d'estimation des jambes par VIBE due à l'occlusion de celles-ci par les cordes, tandis que PARE n'en est pas affecté (Figure 3.6).

Figure 3.5 Estimation de pose par VIBE

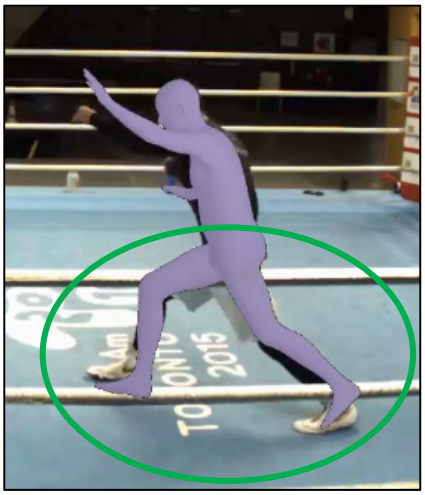


Figure 3.6 Estimation de pose par PARE

Les poses tridimensionnelles produites sont inspectées visuellement une à une par l'auteure de ce mémoire afin de vérifier qu'elles correspondent bien à la pose réelle des boxeurs sur les vidéos.

3.2 Classificateur pour la reconnaissance d'actions en boxe

Une fois constituée, la base de données de poses 3D sert à l'entraînement et l'évaluation d'un classificateur pour la reconnaissance d'actions en boxe. Les détails de celui-ci sont présentés dans cette section.

3.2.1 Description du modèle et des hyperparamètres d'entraînement

Le schéma-bloc du réseau de neurones entraîné pour la reconnaissance d'actions en boxe à partir des poses est illustré à la figure 3.7. Les hyperparamètres utilisés pour l'entraînement du réseau sont indiqués dans le tableau 3.5. Ceux-ci ont fait l'objet d'une recherche en grille afin d'en déterminer les valeurs optimales. Les résultats détaillés se trouvent en ANNEXE V.

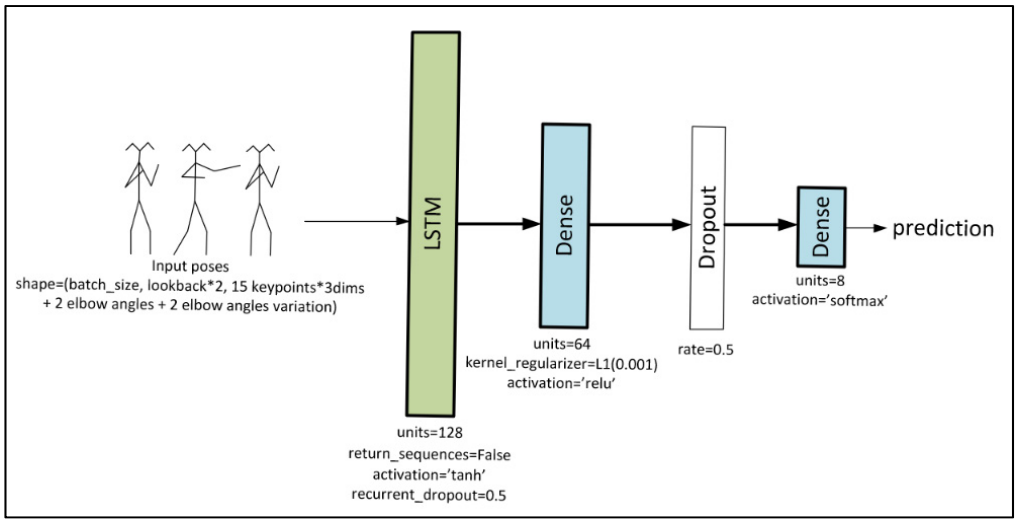


Figure 3.7 Architecture du réseau de neurones

Tableau 3.5 Hyperparamètres d'entraînement

Optimiseur	Type	SGD
	Taux d'apprentissage initial	0,01
	Momentum	0,95
	Nesterov	Oui
Réduction du taux d'apprentissage	Facteur de réduction du taux d'apprentissage	0,1
	Seuil minimal du taux d'apprentissage	0,001
	Valeur suivie	perte
	Changement minimal	0,01
	Patience	2
Arrêt prématuré	Valeur suivie	perte
	Changement minimal	0,01
	Patience	5
Taille d'un lot		32
Longueur de la série temporelle		14

Le classificateur utilisé dans le cadre de cette maîtrise est de type LSTM. Bien que la revue de la littérature ait mis en lumière les meilleures performances des GCN pour la

reconnaissance d'actions à partir de poses, des expérimentations ont permis de constater que pour les données spécifiques à ce projet, un LSTM se révèle plus performant. Le réseau LSTM présenté ci-haut est comparé à un GCN faisant partie de l'état des connaissances, soit Shift-GCN (Cheng *et al*., 2020), présenté précédemment dans la revue de la littérature à la section 1.4.2. L'exactitude et le score F1 sont tous deux plus élevés pour le LSTM (Tableau 3.6).

Tableau 3.6 Comparaison LSTM vs GCN sur les données spécifiques à la boxe

Métrique	LSTM	GCN
Exactitude (↑)	**0,79**	0,71
Score F1 macro (↑)	**0,67**	0,46

La base de données spécifiques à la boxe contient moins d'actions que l'ensemble de référence NTU RGB+D. Toutefois, celles-ci partagent davantage de similarités, puisqu'il s'agit principalement de mouvements rapides des membres supérieurs. Ainsi, il est possible que la temporalité des données soit plus utile pour discriminer les différentes positions et techniques offensives en boxe que la spatialité, qui relève plutôt des GCN. Tel que mentionné dans la revue de la littérature, les LSTM ont été élaborés spécifiquement pour traiter des séries temporelles, ce qui en fait un choix intuitif pour ce projet.

3.2.2 Préparation des données

Les données brutes issues de l'estimation de pose consistent en 75 valeurs, qui représentent les coordonnées x, y, z de 25 parties du corps humain. Seules 15 parties du corps (Figure 3.8) sont retenues pour l'entraînement du modèle, puisque la recherche en grille présentée à l'ANNEXE V semble indiquer que les autres n'ont pas une grande contribution pour discriminer les positions et techniques offensives en boxe. De plus, leurs vitesses en x, y, z sont calculées, en plus des angles des coudes et de leur variation entre deux poses consécutives afin d'augmenter les caractéristiques fournies en entrée au réseau. Un exemple est donc constitué au final de 94 caractéristiques (15 parties du corps * 3 dimensions + 15

parties du corps * 3 vitesses linéaires + 2 angles des coudes gauche et droit + 2 variations des angles des coudes gauche et droit).

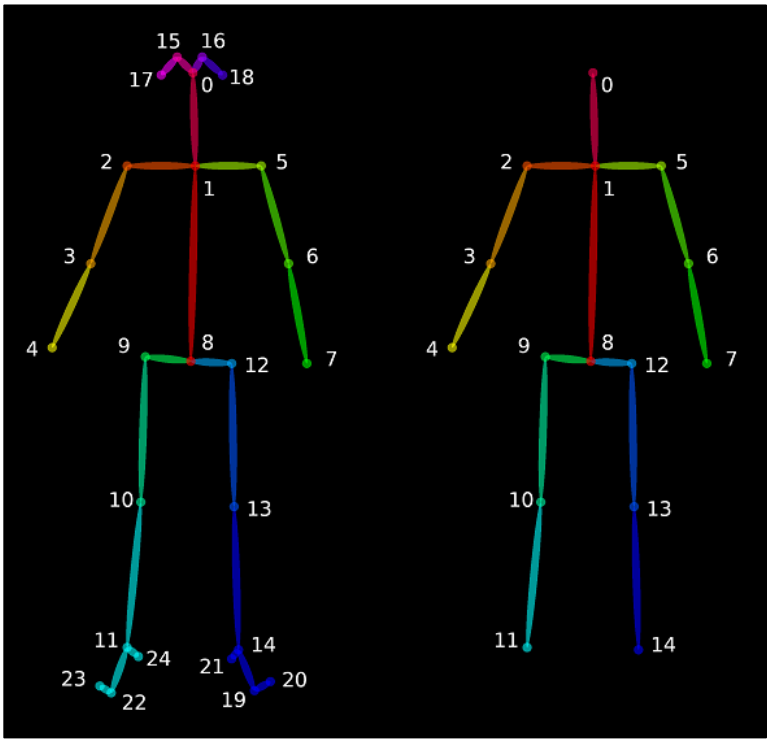


Figure 3.8 Parties du corps d'origine (gauche) et parties du corps retenues (droite)

Les caractéristiques résultantes doivent être regroupées en séries temporelles afin de pouvoir être utilisées pour entraîner un réseau LSTM. Typiquement, le réseau apprend à classifier le dernier élément d'une série temporelle (Figure 3.9).

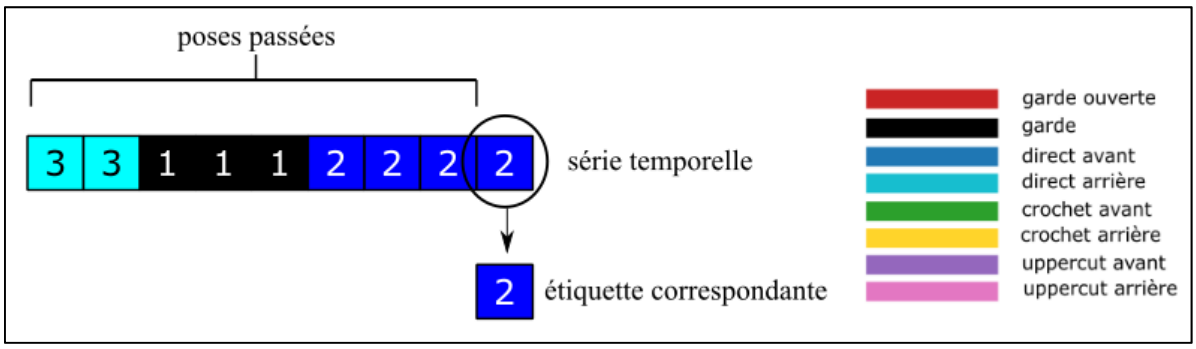


Figure 3.9 Composition typique d'une série temporelle et étiquette correspondante

Puisque dans le cadre de ce projet de maîtrise, la reconnaissance d'actions n'a pas besoin d'être réalisée en temps réel, il est intéressant de tirer profit du fait de détenir à la fois de l'information passée et future. Ainsi, l'étiquette attribuée à la série temporelle est celle de la pose centrale (Figure 3.10).

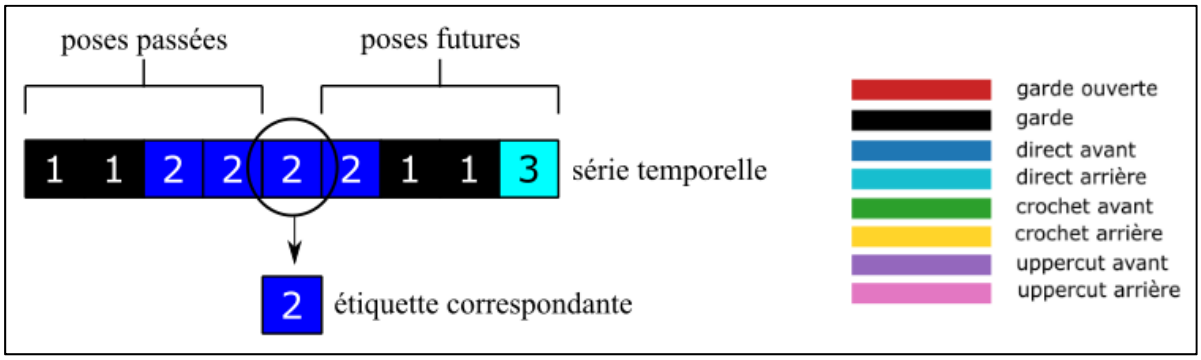


Figure 3.10 Composition des séries temporelles et étiquette correspondante

3.2.3 Augmentation de données

L'entraînement d'un réseau de neurones requiert une énorme quantité de données. Dans ce projet, ce nombre est limité d'une part par la disponibilité de vidéos de *shadow boxing* sur la plateforme YouTube et celles fournies par l'INS Québec, et d'autre part par le temps d'étiquetage desdites vidéos dans le temps alloué pour compléter cette présente maîtrise. Une pratique courante pour obtenir un grand ensemble de données d'entraînement à partir d'un nombre limité de celles-ci est l'augmentation de données. Les quatre types d'augmentations réalisées sur les poses dans le cadre de cette maîtrise sont l'inversion horizontale, la rotation, l'augmentation temporelle et l'ajout de bruit. Elles sont effectuées sur l'ensemble d'entraînement.

L'inversion horizontale permet de simuler des poses où la main dominante du boxeur est celle inverse des poses d'origine (Figure 3.11 et Figure 3.12).

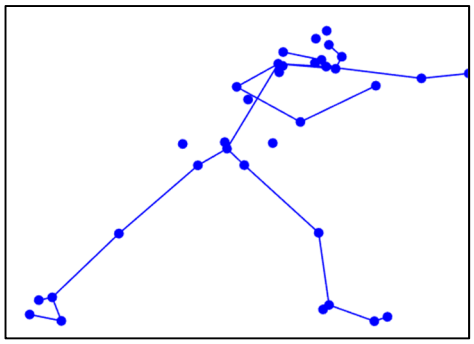


Figure 3.11 Pose sans inversion

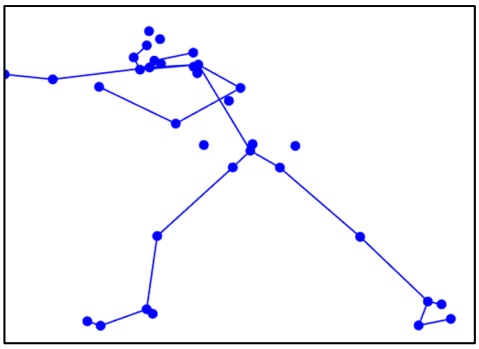


Figure 3.12 Pose inversée

La rotation permet quant à elle de simuler d'autres points de vue de la caméra (Figure 3.13 et Figure 3.14). Dans le cadre de ce projet de maîtrise, les poses sont tournées autour de l'axe longitudinal avec des angles entre 5 et 355 degrés, par incrément de 5.

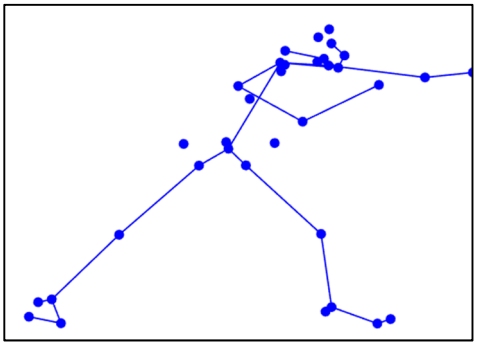


Figure 3.13 Pose sans rotation

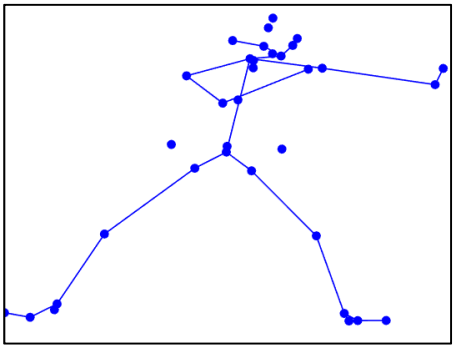


Figure 3.14 Pose tournée à 25°

L'augmentation de données temporelle consiste à simuler une vitesse plus lente ou plus rapide des actions. Les facteurs multiplicatifs utilisés sont 0,75 et 1,25. Pour le premier cas, il suffit de sous-échantillonner les données. Le second cas requiert une interpolation pour générer plus de données et ainsi, obtenir une version ralentie du mouvement.

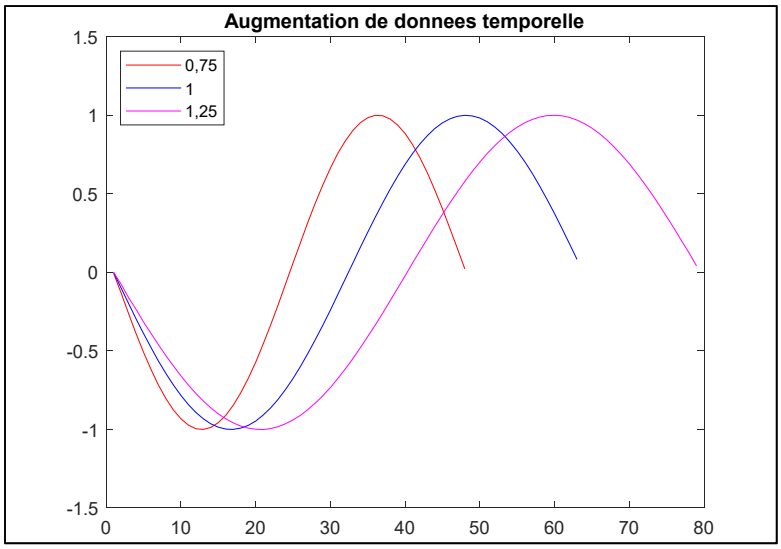


Figure 3.15 Signal ralenti et accéléré par l'augmentation de données temporelle

Finalement, l'ajout de bruit est une technique répandue d'augmentation de données qui permet de modifier légèrement les données d'entrée afin de rendre le réseau robuste à ces légères variations du mouvement. Dans le cadre de ce projet de maîtrise, du bruit gaussien de moyenne 0 et d'écart-type 0,01 est appliqué aux coordonnées tridimensionnelles.

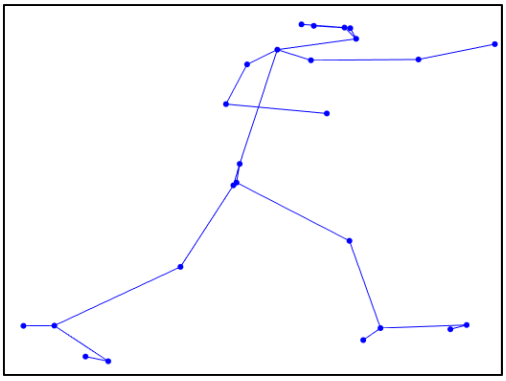


Figure 3.16 Pose sans bruit

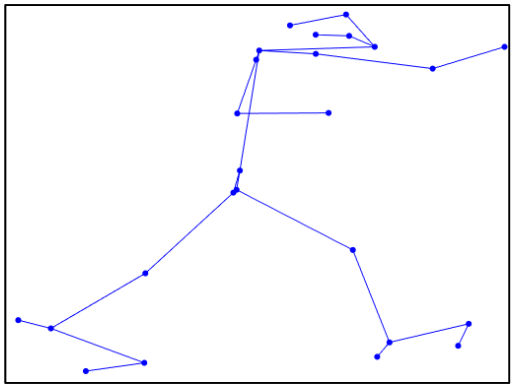


Figure 3.17 Pose bruitée

3.2.4 Segmentation par action et filtrage des prédictions

Lors de l'évaluation classique du modèle, ce dernier assigne une étiquette à chaque pose de la vidéo d'entrée, puis calcule les métriques de classification pose par pose. Ceci engendre deux

problèmes qui réduisent les performances perçues du classificateur. D'une part, un même mouvement peut résulter en la prédiction de deux actions distinctes (Figure 3.18), ce qui amène des faux positifs. D'autre part, si le modèle ne détecte pas l'initiation du mouvement et son achèvement aux moments précis où ceux-ci surviennent (Figure 3.19), les performances de classification en sont réduites puisque certaines poses seront alors considérées comme des faux négatifs ou des faux positifs. Dans ces deux cas de figure, les performances calculées pose par pose ne reflètent pas que le classificateur a bel et bien détecté l'action adéquate.

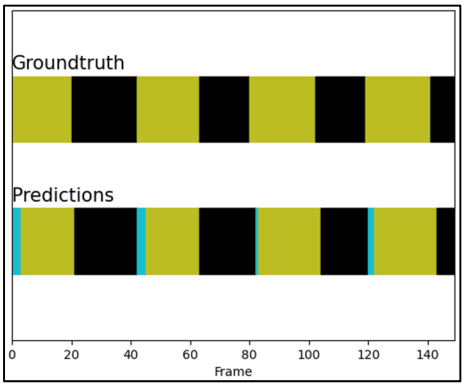


Figure 3.18 Détection de deux actions pour un mouvement

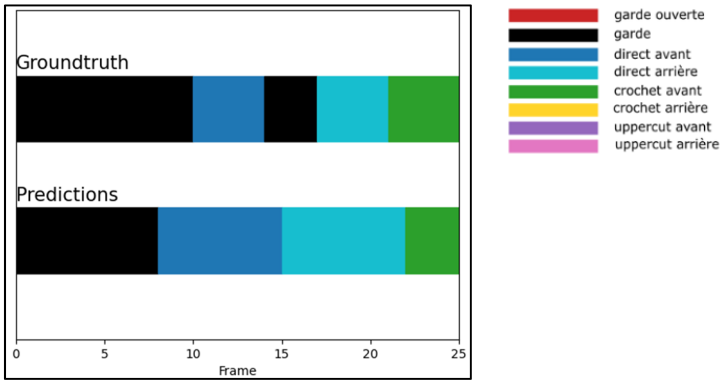


Figure 3.19 Mauvaise estimation du début et de la fin de l'action

Les métriques doivent donc plutôt être calculées par action. Chaque suite d'une même étiquette est considérée comme une seule action. Puis, pour détecter les vrais et les faux positifs, les actions prédites sont comparées aux étiquettes de la vérité terrain contenues entre les mêmes bornes temporelles (Figure 3.20). Un vrai positif est détecté lorsqu'une action prédite se trouve au moins une fois parmi la vérité terrain, dans la même période de temps. À l'inverse, un faux positif survient lorsqu'une action prédite n'apparaît aucune fois dans la vérité terrain entre les mêmes bornes temporelles.

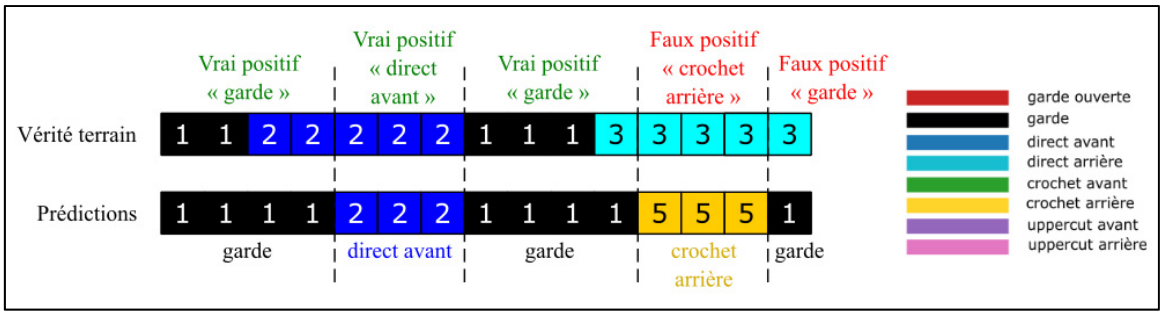


Figure 3.20 Détection des vrais et faux positifs

De manière similaire, si l'action de la vérité terrain ne se trouve aucune fois dans les prédictions entre les mêmes bornes, il s'agit d'un faux négatif.

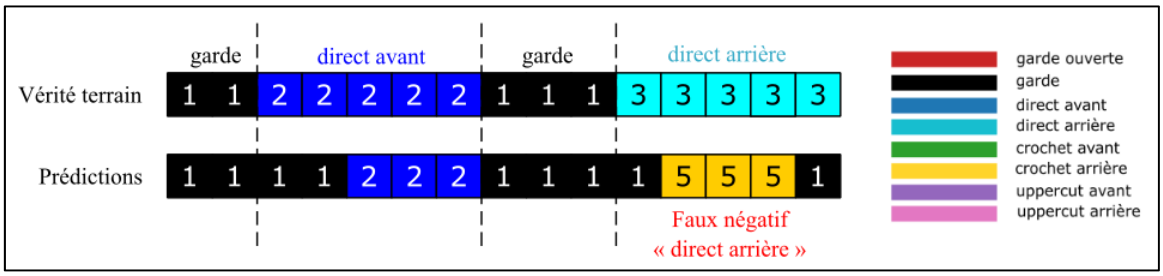


Figure 3.21 Détection des faux négatifs

De plus, les prédictions sont préalablement filtrées pour faciliter la segmentation par action par les règles empiriques de haut niveau suivantes :

1) Les prédictions consécutives de la classe « garde ouverte » doivent avoir une confiance de prédiction médiane supérieure à 90 %;
2) Les poses prédites comme des coups de poing doivent avoir une confiance de prédiction supérieure à 85 %, sinon elles sont converties en l'action voisine ayant la plus haute confiance de prédiction;
3) Si différentes actions de la même main sont prédites de manière consécutive, l'ensemble est converti en l'action la plus longue.

3.3 Algorithme de suivi des boxeurs

Afin de pouvoir suivre les déplacements des athlètes sur l'aire de combat, un algorithme de suivi est développé. Les étapes pour y arriver sont schématisées dans la figure 3.22 et sont plus longuement expliquées dans cette section.

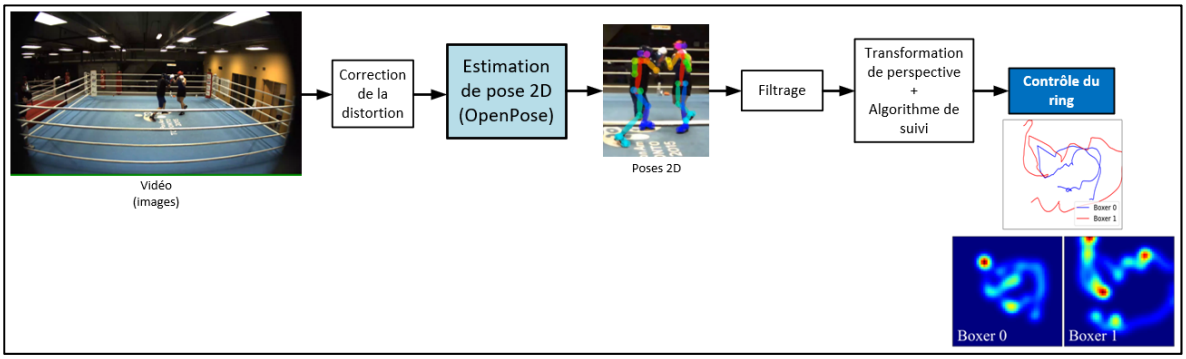


Figure 3.22 Étapes de l'analyse des déplacements des boxeurs

3.3.1 Description des vidéos et prétraitement des images

Contrairement à la reconnaissance d'actions, le suivi des boxeurs ne se limite pas aux vidéos de *shadow boxing*. Ainsi, des images de combats entre deux athlètes peuvent être prises en charge. Celles-ci sont fournies par l'INS Québec. De résolution 1920 par 1080 pixels, elles contiennent l'entièreté de l'aire de combat et ont un point de vue fixe. Le logiciel Corel PaintShop Pro est utilisé pour corriger l'effet de distorsion de la lentille pouvant être observé sur les vidéos d'entrée (Figure 3.23 et Figure 3.24).

Figure 3.23 Image d'origine

Figure 3.24 Correction de la distorsion

3.3.2 Estimation de pose 2D

Les estimateurs de pose 3D utilisés pour constituer la base de données décrite dans la section 3.1 de ce chapitre ne donnent pas les poses dans un système de coordonnées global, mais plutôt dans un référentiel où l'origine est en tout temps située à la mi-hanche de la personne détectée. VIBE et PARE ne peuvent donc pas être utilisés tels quels pour le suivi des boxeurs dans des vidéos de combat.

Pour cette partie, l'estimateur de pose utilisé est OpenPose (Cao *et al*., 2019), qui prédit plutôt les poses 2D dans une image. Celles-ci représentent les pixels associés aux points-clés du corps humain, ce qui permet d'obtenir un système de coordonnées global.

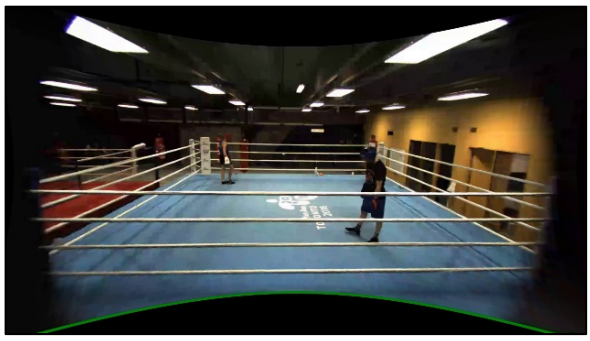


Figure 3.25 Image d'entrée

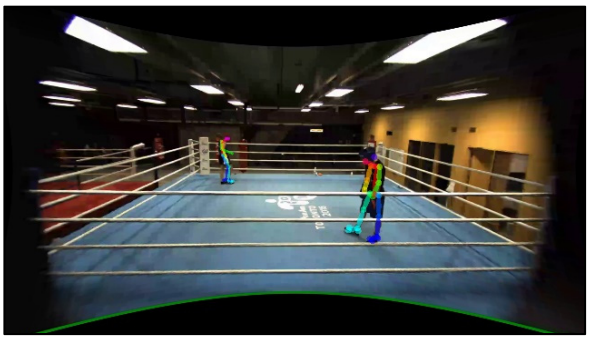


Figure 3.26 Résultat OpenPose

3.3.3 Filtrage

Les détections produites par OpenPose se révèlent erronées à certains moments. Les trois cas de figure se produisant dans les vidéos de combat sont les fausses détections, la détection d'une personne ne participant pas au combat et l'occlusion entre les boxeurs. L'algorithme de suivi est rédigé de manière à être robuste malgré ces erreurs grâce à un algorithme de filtrage (Algorithme 1). Il est tenu pour acquis qu'il y aura toujours deux boxeurs sur l'aire de combat, et qu'OpenPose les détecte adéquatement sur la première image.

Algorithme 1 Filtrage des positions des boxeurs

Données : positions ← positions du point-milieu entre les chevilles des deux boxeurs
seuil ← 0,5

```
for pose i = 1, 2, …, total images dans la vidéo
    if i = 1 (première image)
        Assigner les premières détections aux positions actuelles et précédentes des deux boxeurs
        position_actuelle_boxeur(i, 1) ← détections(i, 1)
        position_actuelle_boxeur(i, 2) ← détections(i, 2)
        position_précédente_boxeur(1) ← détections(i, 1)
        position_précédente_boxeur(2) ← détections(i, 2)
    else (images suivantes)
        id_plus_proche_personne = [3, 3]    3 est la valeur arbitraire indiquant une fausse détection

        if une seule personne détectée (détection(i, 2) n'existe pas)
            Calculer la distance euclidienne entre la détection 1 et les positions précédentes
            dist_euclid(1) = distance_euclidienne(position_précédente_boxeur(1), détection(i, 1))
            dist_euclid(2) = distance_euclidienne(position_précédente_boxeur(2), détection(i, 1))

            Déterminer l'ID de la plus petite distance euclidienne
            id_plus_proche_position = argmin(dist_euclid)
                Assigner la détection à la bonne personne si le déplacement n'est pas anormalement élevé
                if dist_euclid(plus_proche_position) < seuil_hauteur · hauteur_corps
                    id_plus_proche_personne(id_plus_proche_personne) = 1

        else if deux personnes détectées
            Calculer les différentes distances euclidiennes entre les détections actuelles et les positions
            précédentes
            dist_euclid(1, 1) = distance_euclidienne(position_précédente_boxeur(1), détection(i, 1))
            dist_euclid(1, 2) = distance_euclidienne(position_précédente_boxeur(1), détection(i, 2))
            dist_euclid(2, 1) = distance_euclidienne(position_précédente_boxeur(2), détection(i, 1))
            dist_euclid(2, 2) = distance_euclidienne(position_précédente_boxeur(2), détection(i, 2))

            Calculer les ID de la plus petite distance euclidienne de chaque personne
            id_plus_petite_dist(1) = argmin(dist_euclid(1, :))
            id_plus_petite_dist(2) = argmin(dist_euclid(2, :))

            Calculer l'ID de la plus petite distance parmi les plus petites distances de chaque personne
            id_plus_proche_personne = min(dist_euclid(1, id_plus_petite_dist(1)), dist_euclid(2,
```

```
                id_plus_petite_dist(2)))

                    Assigner la détection la plus proche à la bonne personne si le déplacement n'est pas
                    anormalement élevé
                    if dist_euclid(id_plus_proche_personne) < seuil_hauteur · hauteur_corps
                        position_actuelle_boxeur(i, id_plus_proche_personne) ←
                            détections(i, id_plus_proche_personne)
                        Assigner la seconde distance la plus proche à l'autre personne si le déplacement n'est pas
                        anormalement élevé
                        if dist_euclid(id_plus_proche_personne-1) < seuil_hauteur · hauteur_corps
                            position_actuelle_boxeur(i, id_plus_proche_personne-1) ←
                            détections(i, id_plus_proche_personne-1)

                    for personne j = 1, 2
                        if id_plus_proche_personne(j) = 3        Fausse détection ou détection manquante
                            Assigner la position précédente à la position actuelle
                            position_actuelle_boxeur(i, j) ← position_précédente_boxeur(i, j)
                        else
                            Assigner la détection la plus proche
                            position_actuelle_boxeur(i, j) ← détections(i, id_plus_proche_personne)
```

Sur la figure 3.27, il est possible de voir qu'OpenPose confond un sac de frappe avec une personne. L'algorithme de filtrage permet toutefois d'ignorer cette fausse détection et de conserver en mémoire les positions des deux boxeurs.

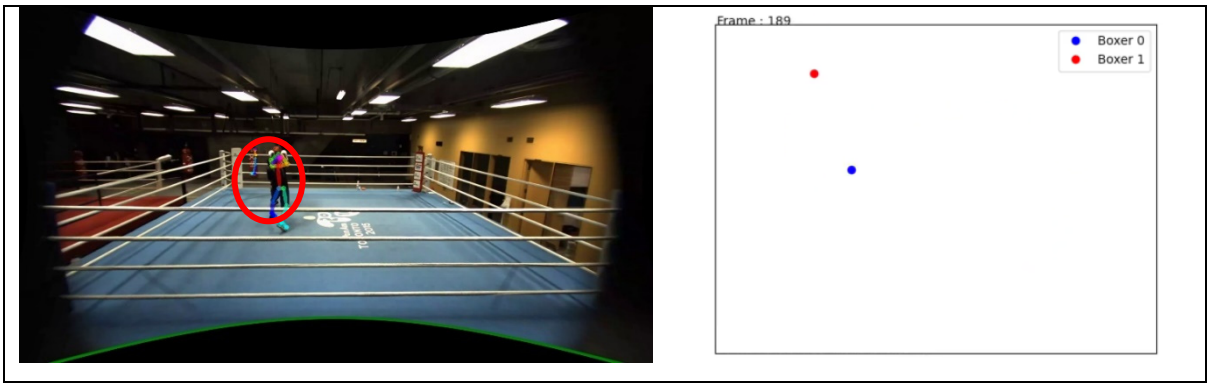


Figure 3.27 Résultat du filtrage sur une fausse détection OpenPose

D'autres personnes que les duellistes peuvent être présentes sur les vidéos de combat fournies par l'INS Québec. Par exemple, l'entraîneur se trouve souvent près du ring pour

s'assurer du bon déroulement de la séance (Figure 3.28). D'autres athlètes de l'Institut peuvent également circuler en arrière-plan. Encore une fois, l'algorithme de filtrage permet d'ignorer toute autre personne que les deux boxeurs.

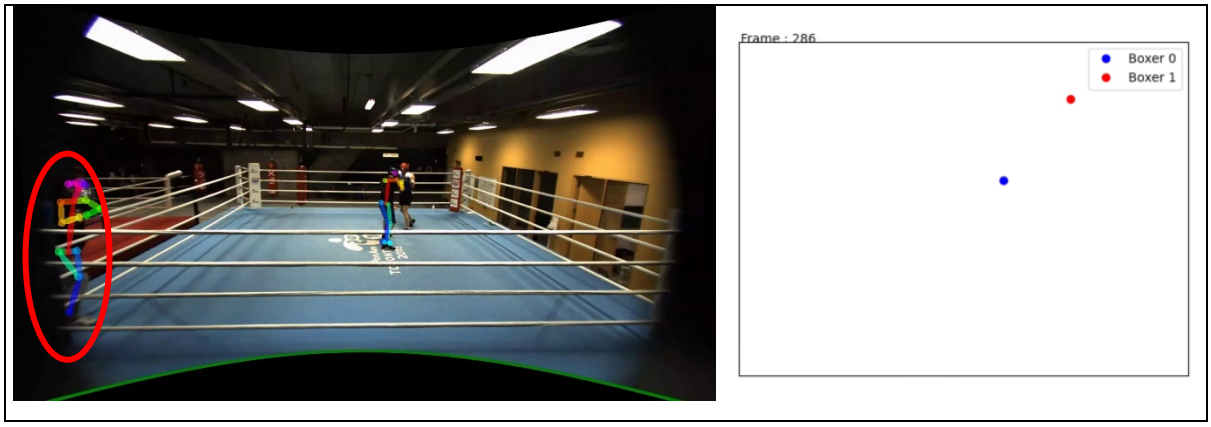


Figure 3.28 Résultat du filtrage sur une autre détection OpenPose

Finalement, lorsqu'un boxeur se trouve derrière l'autre sur la vidéo, OpenPose n'arrive parfois qu'à détecter l'athlète le plus près de la caméra (Figure 3.29). Il s'agit du problème d'occlusion, bien connu en vision par ordinateur. En gardant en mémoire la dernière position du boxeur non détecté, l'algorithme de filtrage permet d'afficher l'emplacement des boxeurs en continu malgré les possibles occlusions.

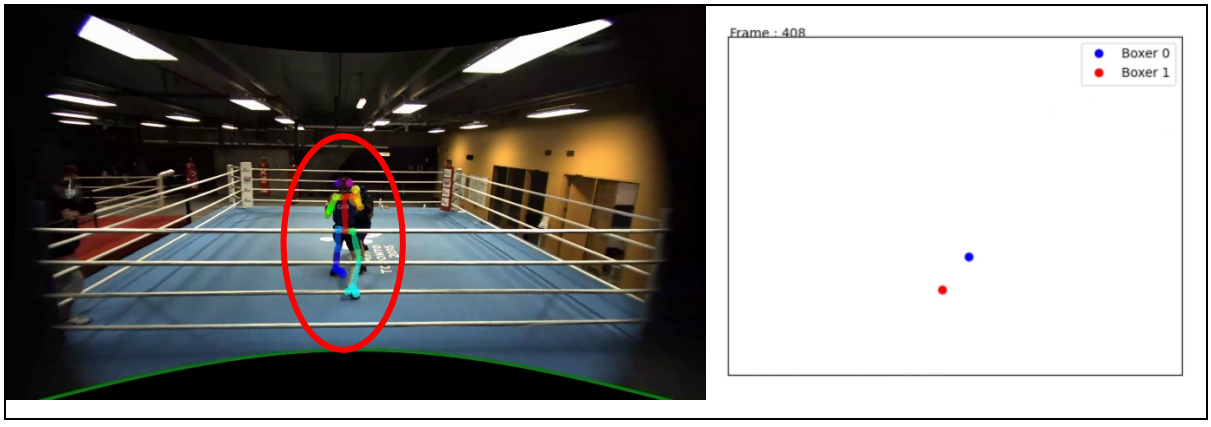


Figure 3.29 Résultat du filtrage sur un cas d'occlusion

3.3.4 Transformation de perspective

Le point-milieu entre les chevilles détectées par OpenPose constitue le repère correspondant à la position du boxeur sur l'aire de combat. Ce choix est fait afin d'obtenir une bonne estimation du centre du corps de l'athlète projeté au niveau de la surface de combat.

Les variables requises pour la transformation de perspective sont décrites dans le tableau 3.7 et illustrées sur la figure 3.30. En connaissant les coordonnées en pixels de ce repère, des quatre coins de l'aire de combat ainsi que les dimensions réelles de cette dernière, il est possible de convertir la position du boxeur du domaine de l'image vers celui du ring.

Tableau 3.7 Variables requises pour la transformation de perspective

Variable	Description
(x_img, y_img)	coordonnées en pixels du repère sur l'image
(x1_img, y1_img) *(x2_img, y2_img)* *(x3_img, y3_img)* *(x4_img, y4_img)*	coordonnées en pixels des quatre coins de l'aire de combat sur l'image
ring_dim	dimension réelle du côté de l'aire de combat

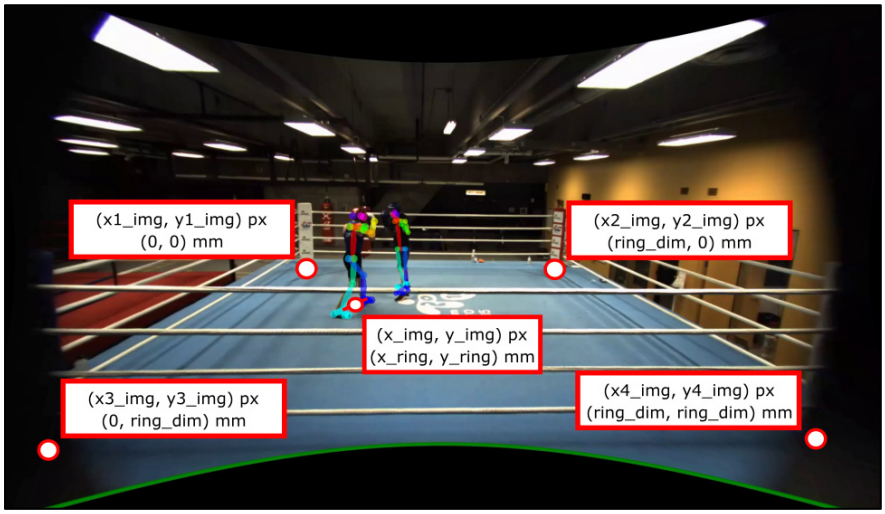


Figure 3.30 Repères pour la transformation de perspective

Dans un premier temps, les coordonnées en pixels des quatre coins de l'aire de combat sont inscrites dans un vecteur de dimensions 4x2 :

$$ring\_img = ([x1\_img, y1\_img], [x2\_img, y2\_img], [x3\_img, y3\_img], [x4\_img, y4\_img]) \tag{3.1}$$

Les coordonnées réelles sont quant à elle inscrites dans un second vecteur de dimensions 4x2 :

$$ring\_réel = ([0,0], [ring\_dim, 0], [0, ring\_dim], [ring\_dim, ring\_dim]) \tag{3.2}$$

La matrice de transformation de perspective *M* est calculée grâce à la fonction *getPerspectiveTransform* de la librairie OpenCV :

$$M = cv2.getPerspectiveTransform(ring\_img, ring\_réel) = \begin{bmatrix} M_1 & M_2 & M_3 \\ M_4 & M_5 & M_6 \\ M_7 & M_8 & M_9 \end{bmatrix} \tag{3.3}$$

Le produit matriciel de *M* et du vecteur augmenté des coordonnées en pixels du repère situé à la mi-cheville du boxeur permet de calculer des valeurs *a*, *b* et *c*. Les valeurs *a* et *b* sont en fait les coordonnées *x_réel* et *y_réel* multipliées par un facteur d'échelle *s* (ou *c*) permettant de passer du référentiel de l'image à celui de l'aire de combat.

$$M \cdot \begin{bmatrix} x\_img \\ y\_img \\ 1 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} s \cdot x\_réel \\ s \cdot y\_réel \\ s \end{bmatrix} \tag{3.4}$$

Pour obtenir la position du boxeur dans le système de coordonnées de l'aire de combat, il ne suffit donc que de diviser les valeurs *a* et *b* par la valeur *c* :

$$[x\_réel \quad y\_réel] = \begin{bmatrix} \frac{a}{c} & \frac{b}{c} \end{bmatrix} \tag{3.5}$$

3.3.5 Validation de l'algorithme

Afin d'évaluer la précision de l'algorithme de suivi des boxeurs, 16 repères circulaires de 13,5 po de diamètre sont disposés à des distances connues sur l'aire de combat (Figure 3.31).

Afin que leur centre soit bien visible sur les images, celui-ci est marqué d'un cercle noir plein de 4 po de diamètre. Puis, la vidéo d'une personne parcourant la surface (Figure 3.32) tout en posant ses pieds sur la circonférence de tous les disques (Figure 3.33) est enregistrée. En comparant la position du point-milieu des pieds obtenus par l'algorithme aux positions réelles connues du centre des repères, il est ainsi possible de calculer l'erreur moyenne entre ces valeurs. Celle-ci correspond à la distance euclidienne moyenne entre les positions réelles et les positions calculées par l'algorithme :

$$Erreur = \frac{\sum_{i=1}^{16} \sqrt{\left(x_{i,reel} - x_{i,calc}\right)^2 + \left(y_{i,reel} - y_{i,calc}\right)^2}}{16} \tag{3.6}$$

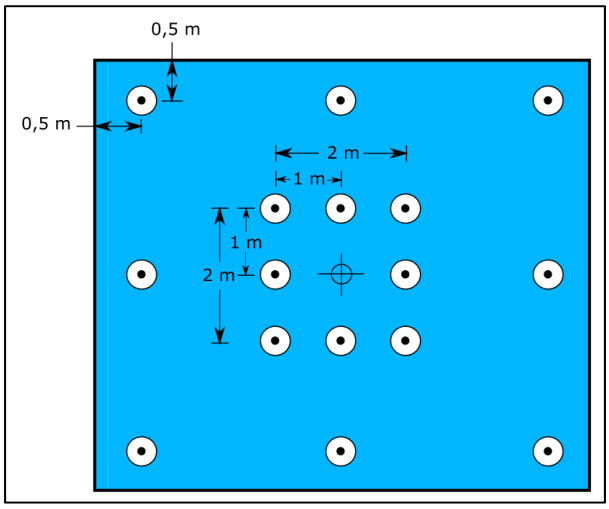


Figure 3.31 Distance des repères

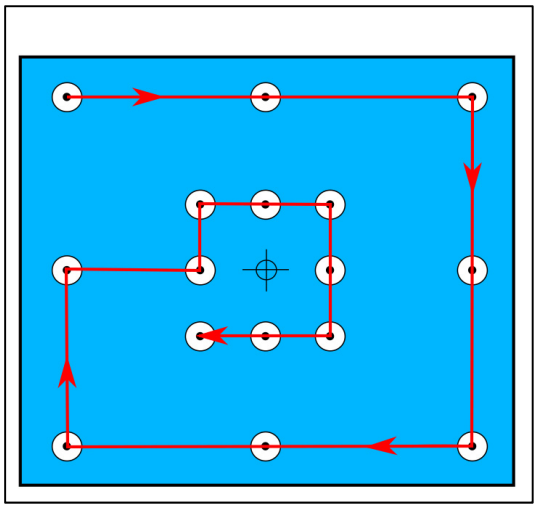


Figure 3.32 Exemple de parcours

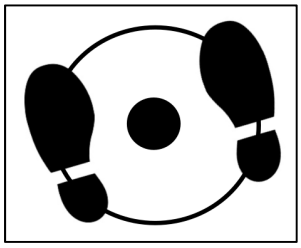


Figure 3.33 Position des pieds sur le repère

Afin de vérifier si l'orientation de la personne par rapport à la caméra a une influence sur l'erreur, celle-ci pose ses pieds autour de chaque repère à deux reprises, en faisant face à un

endroit différent pour chacune d'elle. De plus, pour une plus grande validité interne, l'expérimentation décrite ci-dessus est réalisée deux fois, chacune par une personne différente.

3.4 Extraction des métriques de performance en boxe

Le classificateur et l'algorithme de suivi des boxeurs décrits respectivement à la section 3.2 et à la section 3.3 permettent ainsi l'extraction de métriques de performance en boxe. Celles-ci sont détaillées dans cette section.

Dans un premier temps, un compteur d'action est mis en place grâce aux prédictions effectuées par le classificateur. Il indique le nombre de coups offensifs donnés, parmi les coups droits avant et arrière, les crochets avant et arrière, ainsi que les uppercuts avant et arrière. Puis, il calcule le temps passé en position de garde ouverte et de garde fermée. Comme mentionné précédemment, chaque séquence d'une même étiquette consécutive est considérée comme une seule action. Il ne suffit donc que de tenir le compte de ces groupes d'étiquettes pour obtenir le nombre d'occurrences de chaque action. Pour calculer le temps passé en position de garde ouverte ou de garde fermée, le nombre de poses contenues dans le bloc identifié comme une action de ce type est divisé par la fréquence d'images de la vidéo d'entrée. Finalement, pour obtenir une mesure de l'occupation de l'aire de combat par chaque boxeur, l'algorithme de suivi est utilisé afin de tracer une carte des déplacements totaux des athlètes. Des cartes de chaleur de la position des boxeurs peuvent également être générées pour quantifier le temps passé à certains endroits de l'aire de combat.

CHAPITRE 4

ÉTUDE COMPARATIVE

Une étude comparative est effectuée afin d'évaluer la pertinence de sélectionner ou non certains éléments du réseau de neurones, de son entraînement et de ses données d'entrée. Les métriques comparées sont la précision, le rappel et le score F1. La précision est le rapport entre le nombre total de vrais positifs et la somme de tous les positifs (Équation 1). Dans le contexte de la reconnaissance d'actions, il s'agit du pourcentage de certitude que lorsqu'une certaine action est détectée, il s'agisse bien de celle indiquée par le réseau de neurones.

$$Précision = \frac{Vrais\ positifs}{Vrais\ positifs + Faux\ positifs} \tag{4.1}$$

Le rappel est plutôt le rapport entre le nombre total de vrais positifs, et la somme des vrais positifs et des faux négatifs (Équation 2). Il représente le pourcentage des actions qui sont détectées parmi toutes les réelles occurrences de celles-ci.

$$Rappel = \frac{Vrais\ positifs}{Vrais\ positifs + Faux\ négatifs} \tag{4.2}$$

Le score F1, quant à lui, est la moyenne harmonique de la précision et du rappel (Équation 3).

$$Score\ F1 = \frac{2 \cdot Précision \cdot Rappel}{Précision + Rappel} \tag{4.3}$$

L'étude comparative vise à observer l'effet sur les performances de classification de :

1) Différentes stratégies d'augmentation de données;
2) Différentes architectures LSTM;
3) Différentes longueurs des séries temporelles;
4) Différentes stratégies d'étiquetage des données.

Elle permet ainsi de sélectionner les paramètres optimaux pour obtenir un classificateur ayant les plus hautes performances possible.

4.1 Stratégie d'augmentation de données

Les résultats à la suite de différentes méthodes d'augmentation de données discutées à la section 3.2.3 sont présentés dans le tableau 4.1. L'inversion et la rotation sont testées individuellement et de manière combinée, puis les différents facteurs d'augmentation temporelle sont évalués de la même façon sur le meilleur modèle. Finalement, du bruit gaussien est ajouté aux meilleures méthodes d'augmentation de données déterminées jusqu'à ce point.

L'inversion diminue les performances de classification. Cela peut être dû au fait qu'elle ajoute des exemples fausse-patte en entraînement, alors que l'ensemble d'évaluation est majoritairement composé de boxeurs ayant une garde conventionnelle. La rotation augmente quant à elle les métriques de classification. En effectuant une rotation de 5 à 355 degrés par incrément de 5, le nombre de données est multiplié par un facteur 70, en plus d'exposer le classificateur à des points de vue plus variés. La combinaison de l'inversion et la rotation n'améliore pas les métriques de classification par rapport à la rotation seule. Cela est logique puisque l'inversion a un effet nuisible sur les performances de classification. Bien que l'augmentation de données temporelle permette d'améliorer ces dernières sur certaines classes, les performances globales sont inférieures à la rotation seule. La temporalité des données d'entraînement initiales est donc suffisamment représentative des données d'évaluation. L'ajout de bruit combiné à la rotation n'améliore pas les performances dans l'ensemble. Il est possible que le bruit déforme trop les poses de l'ensemble de données d'entraînement, de sorte que celles-ci ne soient plus représentatives des mouvements naturels retrouvés dans l'ensemble d'évaluation. Ainsi, la rotation est la stratégie d'augmentation des données retenue.

Tableau 4.1 Comparaison de différentes stratégies d'augmentation de données

Classe	Précision (%)							Rappel (%)							Score F1 (%)							Nombre d'images
	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7	
Garde	35	65	80	61	53	**100**	61	14	**82**	50	73	77	73	64	77	68	49	56	76	68	62	**78**	68	23	22
Garde ouverte	98	96	96	97	96	**100**	97	98	80	78	83	**85**	82	67	81	75	88	86	89	**90**	89	80	88	85	2 711
Direct avant	31	45	57	**62**	**62**	52	49	50	69	95	94	92	92	**100**	94	99	43	61	71	**74**	**74**	69	65	67	104
Direct arrière	30	19	**40**	21	45	19	28	34	40	81	87	83	83	88	85	**90**	34	30	55	33	**59**	31	42	49	52
Crochet avant	49	53	53	**60**	**60**	38	57	47	73	57	86	88	85	93	**94**	87	59	55	66	**71**	70	54	**71**	61	153
Crochet arrière	41	39	49	**50**	48	33	43	47	8	86	**94**	78	**94**	85	90	83	56	54	**65**	61	64	48	59	60	314
Uppercut avant	35	32	58	**63**	52	45	47	52	57	52	75	77	81	65	63	**85**	44	39	65	**69**	63	53	54	65	233
Uppercut arrière	26	17	70	**77**	63	60	68	61	9	3	31	**35**	26	27	17	31	14	5	43	**49**	37	37	28	41	347
moyenne sur toutes les classes	43	46	**63**	61	60	56	56	50	62	63	**78**	77	77	74	75	77	48	48	**66**	64	65	56	59	59	-
moyenne sans « garde » et « garde ouverte »	35	34	55	**56**	53	41	49	49	56	62	78	76	77	74	74	**79**	42	41	**61**	60	**61**	49	53	57	-

Légende des stratégies d'augmentation de données :

0 : aucune
1 : inversion
2 : rotation
3 : inversion + rotation
4 : augmentation temporelle (accéléré, facteur 0,75)
5 : augmentation temporelle (ralenti, facteur 1,25)
6 : augmentation temporelle combinée
7 : ajout de bruit gaussien

4.2 Architecture LSTM

En se servant de la meilleure stratégie d'augmentation de données, les différentes manières de préparer les données pour le réseau LSTM discutées à la section 3.2.2 sont comparées (Tableau 4.2). De plus, un réseau LSTM bidirectionnel (Bi-LSTM) est entraîné afin de comparer ses performances au LSTM standard et au LSTM entraîné avec des poses futures. Le Bi-LSTM lit les séries temporelles dans le sens conventionnel, ainsi que dans la direction inverse. Il s'agit d'une autre façon de bénéficier d'information future en entraînement.

Tableau 4.2 Comparaison de différentes architectures LSTM

Classe	Précision (%)			Rappel (%)			Score F1 (%)			Nombre d'images		
	LSTM	Bi-LSTM	Poses futures	LSTM	Bi-LSTM	Poses futures	LSTM	Bi-LSTM	Poses futures	LSTM	Bi-LSTM	Poses futures
Garde ouverte	39	68	**80**	33	48	**73**	36	57	**76**	27	27	22
Garde	95	**96**	**96**	82	79	**83**	88	87	**89**	2 707	2 707	2 711
Direct avant	43	44	**57**	83	87	**94**	56	58	**71**	95	95	104
Direct arrière	27	37	**40**	**90**	**90**	87	42	52	**55**	60	60	52
Crochet avant	46	36	**53**	85	**86**	**86**	60	51	**66**	143	143	153
Crochet arrière	48	43	**49**	71	80	**94**	57	53	**65**	310	310	314
Uppercut avant	53	45	**58**	**77**	69	75	63	54	**65**	242	242	233
Uppercut arrière	33	31	**70**	10	6	**31**	15	10	**43**	352	352	347

Pour presque toutes les classes, le LSTM recevant à la fois des poses passées et futures présente les meilleures performances. Intuitivement, le réseau détient plus d'information sur le contexte de l'action, ce qui, comme le démontrent les résultats obtenus, bonifie la reconnaissance d'actions. Le LSTM bidirectionnel, qui se base sur la même notion, n'offre toutefois pas d'amélioration significative sur le LSTM standard. Ainsi, la plus-value est obtenue lorsque le LSTM reçoit les poses futures dans le sens dans lequel elles se produisent, tandis que le Bi-LSTM ne fait que lire la séquence d'entrée dans la direction inverse.

4.3 Longueur des séries temporelles

Un paramètre important des réseaux LSTM est la longueur des séries temporelles qu'ils reçoivent en entrée. Dans l'ensemble de données constitué dans le cadre de cette maîtrise, les techniques offensives sont représentées en moyenne par cinq poses. Étant donné que le réseau reçoit en entrée les poses passées et futures, ce nombre est doublé pour créer des séries temporelles de longueur 10. Pour cette portion de l'étude comparative, des tailles inférieures et supérieures sont également employées afin d'évaluer l'influence de la longueur des séquences sur les performances de classification.

Le tableau 4.3 indique que la série temporelle contenant 14 poses permet d'obtenir globalement les meilleures performances de classification. Cela est logique puisque le réseau détient alors plus d'information contextuelle sur l'action à prédire.

Tableau 4.3 Comparaison de différentes longueurs des séries temporelles

Classe	Précision (%)			Rappel (%)			Score F1 (%)			Nombre d'images		
	6	10	14	6	10	14	6	10	14	6	10	14
Garde ouverte	46	**80**	71	54	73	**75**	50	**76**	73	24	22	20
Garde	**98**	96	96	75	83	**88**	85	89	**92**	2 798	2 711	2 644
Direct avant	48	**57**	50	94	94	**99**	64	**71**	66	106	104	100
Direct arrière	**44**	40	36	76	87	**90**	**56**	55	52	59	52	48
Crochet avant	39	53	**64**	83	**86**	82	53	66	**72**	158	153	141
Crochet arrière	40	49	**58**	87	94	**97**	55	65	**72**	314	314	310
Uppercut avant	42	58	**69**	67	**75**	71	51	65	**70**	237	233	231
Uppercut arrière	**76**	70	75	**34**	31	32	**47**	43	45	349	347	342

4.4 Stratégie d'étiquetage

Une problématique de l'étiquetage des données utilisées dans le cadre de cette maîtrise est de déterminer à quels moments précis commencent et terminent les différentes techniques offensives. De plus, l'initiation et la fin de certains coups de poing ont une allure similaire. Différentes stratégies d'étiquetage sont donc comparées afin d'établir la meilleure pour l'entraînement du réseau LSTM (Tableau 4.4).

Tableau 4.4 Comparaison de différentes stratégies d'étiquetage

Classe	Précision (%)				Rappel (%)				Score F1 (%)				Nombre d'images			
	8	9	10	11	8	9	10	11	8	9	10	11	8	9	10	11
Garde ouverte	**100**	70	**100**	71	85	**95**	90	75	92	81	**95**	73	20	20	20	20
Garde	**99**	98	98	96	70	83	81	**88**	82	90	88	**92**	3 353	2 977	3 020	2 644
Direct avant	23	40	31	**50**	**100**	97	97	99	38	57	47	**66**	45	72	73	100
Direct arrière	29	**41**	21	36	**94**	93	92	90	45	**57**	34	52	33	42	39	48
Crochet avant	30	63	55	**64**	81	**93**	48	82	44	**75**	64	72	70	98	113	141
Crochet arrière	19	36	40	**58**	**100**	97	67	97	32	52	50	**72**	104	189	225	310
Uppercut avant	25	50	47	**69**	**87**	73	77	71	39	59	58	**70**	91	171	151	231
Uppercut arrière	43	59	59	**75**	38	26	**43**	32	40	36	**49**	45	120	267	195	342

Légende des stratégies d'augmentation de données :
8 : début et fin exclus du coup de poing (début et fin -> garde)
9 : début inclus dans le coup de poing (fin -> garde)
10 : fin incluse dans le coup de poing (début -> garde)
11 : début et fin inclus dans le coup de poing

Les meilleures performances de classification sont obtenues lorsque l'entièreté du mouvement est considérée comme un coup de poing. Il semble donc que le LSTM bénéficie de l'information contenue dans l'initiation et la fin des mouvements, bien que ceux-ci puissent être similaires d'un coup à l'autre.

L'étude comparative de ce présent chapitre permet donc de déterminer les paramètres optimaux suivants :

1) Augmentation de données par la rotation des poses;
2) Poses futures incluses dans les séries temporelles;
3) Longueur des séries temporelles égale à 14;
4) Début et fin des techniques offensives étiquetés comme coup de poing.

CHAPITRE 5

RÉSULTATS

5.1 Résultats quantitatifs

Cette section présente les résultats de l'évaluation quantitative du classificateur entraîné à reconnaître huit positions et techniques offensives spécifiques à la boxe. De plus, elle contient les résultats de l'expérimentation visant à valider la précision de l'algorithme de suivi des boxeurs.

5.1.1 Performances du classificateur

Afin d'évaluer les performances finales du classificateur avec les paramètres optimaux déterminés par l'étude comparative du chapitre 4 (Tableau 5.1), une validation croisée à 5 plis est effectuée (les résultats complets se trouvent à l'ANNEXE VI). Les totaux des vrais et faux positifs, ainsi que des faux négatifs sont présentés dans le tableau 5.2. Ceux-ci permettent de calculer les valeurs moyennes de la précision, du rappel et du score F1, qui sont présentées dans le tableau 5.3.

Tableau 5.1 Rappel des paramètres optimaux de l'étude comparative

Paramètre	Valeur ou description
Stratégie d'augmentation de données	Rotation entre 5 et 355 degrés, par incrément de 5
Architecture LSTM	Poses futures incluses dans les séries temporelles
Longueur des séries temporelles	14
Stratégie d'étiquetage	Début et fin du mouvement inclus dans le coup de poing

Tableau 5.2 Totaux des vrais positifs, faux positifs et faux négatifs

Classe	Vrais positifs	Faux positifs	Faux négatifs
Garde ouverte	8	9	2
Garde	666	198	36
Direct avant	163	37	18
Direct arrière	156	23	52
Crochet avant	189	50	30
Crochet arrière	43	53	42
Uppercut avant	57	52	59
Uppercut arrière	87	49	41

Tableau 5.3 Rapport de classification multiclasse

Classe	Précision (%)	Rappel (%)	Score F1 (%)	Nombre d'exemples
Garde ouverte	80	47	59	3
Garde	95	77	85	189
Direct avant	90	82	86	40
Direct arrière	75	87	81	38
Crochet avant	86	79	83	48
Crochet arrière	51	45	48	19
Uppercut avant	49	52	51	23
Uppercut arrière	68	64	66	29
MOYENNE	74	67	70	N/A

Le classificateur présente de bonnes performances de classification pour la garde, les directs avant et arrière, ainsi que pour le crochet avant. Le score F1 est supérieur à 80 % pour ces classes. Le modèle est particulièrement faible pour détecter adéquatement les crochets arrière et les uppercuts avant, avec un score F1 près ou inférieur à 50 %.

La matrice de confusion (Figure 5.1) renseigne plus précisément sur les types de techniques offensives les plus souvent confondues entre elles. La figure 5.2 présente quant à elle la

matrice de confusion normalisée selon les colonnes, ou encore, le nombre de positifs de chaque classe, qu'ils soient vrais ou faux.

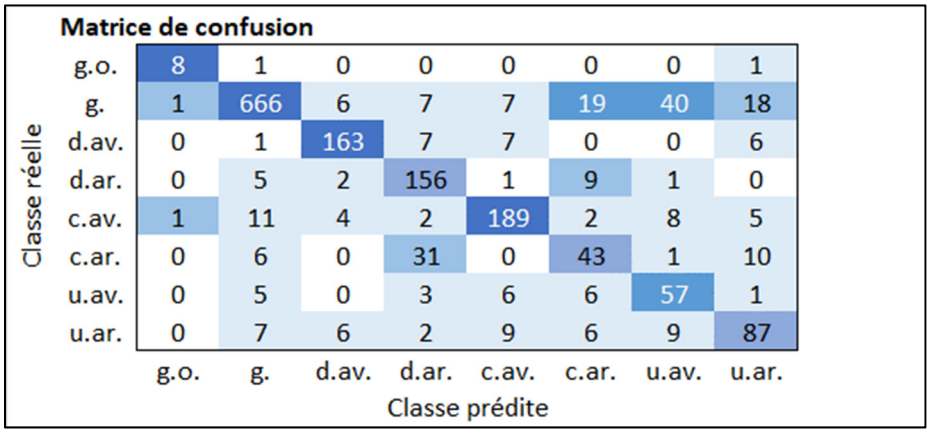


Figure 5.1 Matrice de confusion

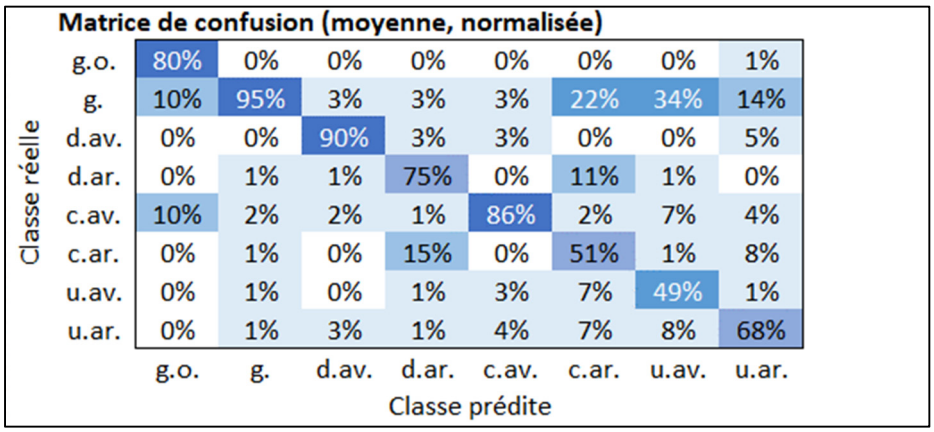


Figure 5.2 Matrice de confusion normalisée

L'exactitude pondérée du classificateur, calculée par l'équation (5.1 à partir de la matrice de confusion, est de 77 %. N est le nombre de classes.

$$Exactitude\ pondérée = \frac{1}{N}\sum_{i=1}^{N}\frac{Vrais\ positifs\ rangée\ i}{Total\ rangée\ i} \tag{5.1}$$

Le tableau 5.4 présente les valeurs de précision, de rappel et de score F1 lorsque seule la détection d'un coup de poing, sans tenir compte de son type, est évaluée. Les valeurs de

87 %, 90 % et 88 % sont beaucoup plus élevées que les moyennes de chaque technique offensive du tableau 5.3, qui sont respectivement en 68 %, 73 % et 67 %. L'exactitude moyenne pondérée est cette fois de 84 %, plutôt que 77 % lorsque le type de coup de poing est pris en compte. Le nombre de coups donnés dans un combat étant un indicateur d'une victoire, le classificateur développé dans le cadre de ce projet est donc à même de renseigner sur cette métrique importante. Il reste toutefois place à amélioration si l'on cherche à déterminer les types de coups donnés afin, par exemple, d'établir les tendances d'un boxeur.

Supposons un combat de trois rondes, dans lesquelles chaque boxeur donne en réalité 50 coups de poing (valeur arrondie d'Ashker, 2011), pour un total de 150. Le classificateur détectera adéquatement l'occurrence de 90 % d'entre eux, soit 135. Il faut cependant s'attendre à avoir la fausse détection de 20 coups de poing excédentaires.

Tableau 5.4 Rapport de classification ternaire

Classe	Précision (%)	Rappel (%)	Score F1 (%)	Nombre d'exemples
Garde ouverte	80	47	59	3
Garde	95	77	85	189
Coup de poing	87	90	88	173
MOYENNE	86	72	76	N/A

L'exactitude moyenne pondérée lorsque l'on s'intéresse au type de coup de poing donné de 77 % est inférieure à la valeur ciblée par l'INS Québec de 80 %. Elle s'en approche toutefois. En ne se souciant pas de distinguer les différentes techniques offensives entre elles, l'exactitude moyenne pondérée est plutôt de 84 %, ce qui est suffisant pour que les experts de l'Institut détiennent de l'information pertinente à l'analyse des performances de leurs athlètes.

5.1.2 Validation de l'algorithme de suivi des boxeurs

Le protocole décrit à la section 3.3.5 mène au calcul d'une erreur moyenne de l'algorithme de suivi de 12,2 cm et d'un écart-type de 5,7 cm (les calculs détaillés sont présentés en ANNEXE VII).

La figure 5.3 présente les valeurs moyennes de l'erreur et son écart-type pour différents emplacements sur l'aire de combat. L'erreur est moins élevée près du coin supérieur gauche, ce qui peut indiquer que les coordonnées en pixels prélevées à cet endroit pour la transformation de perspective sont plus représentatives de leur équivalent réel que les autres coins de l'aire de combat. Outre cette zone, l'erreur demeure malgré tout dans le même ordre de grandeur.

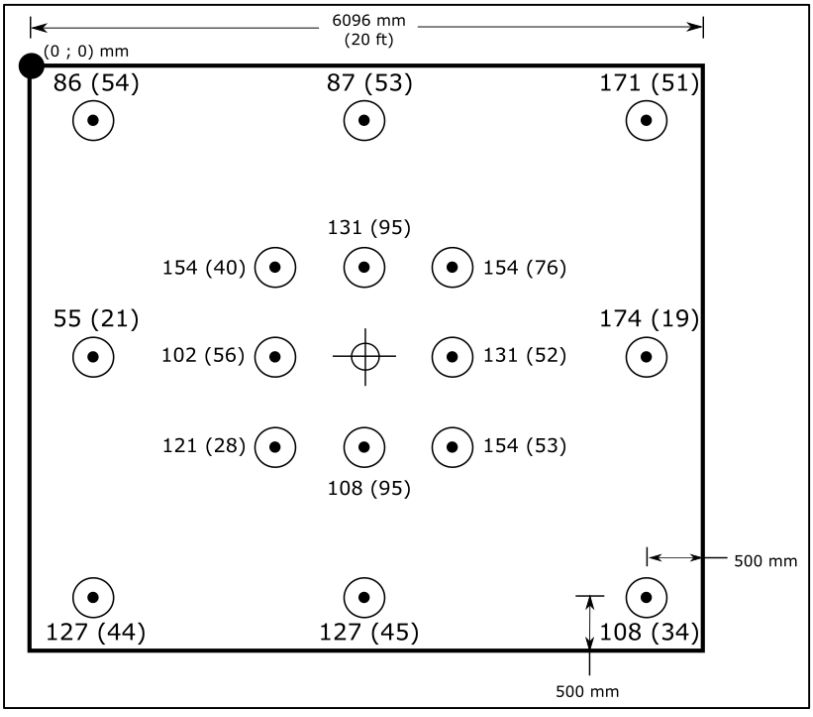


Figure 5.3 Erreur (écart-type) à différents emplacements

De plus, il est possible de comparer l'erreur aux mêmes emplacements, mais selon différentes orientations du corps. La figure 5.4 compare l'erreur obtenue pour le trajet d'une seule personne, aux repères situés au pourtour de l'aire de combat. Les résultats pour la

seconde personne ainsi que pour les repères les plus près du centre du ring sont présentés en ANNEXE VIII. L'orientation du corps par rapport à la caméra semble avoir une influence sur l'erreur, mais le sens de la variation dépend de l'endroit où la personne se situe sur l'aire de combat. Par exemple, une orientation où le plan frontal fait face à la caméra se solde en une erreur plus faible lorsque la personne est au centre des cordes les plus éloignées du point d'observation, tandis qu'elle augmente l'erreur au point-milieu des cordes les plus près de la caméra. Il est ainsi difficile de tirer une conclusion à propos de l'influence de l'orientation du boxeur sur l'erreur de l'algorithme de suivi.

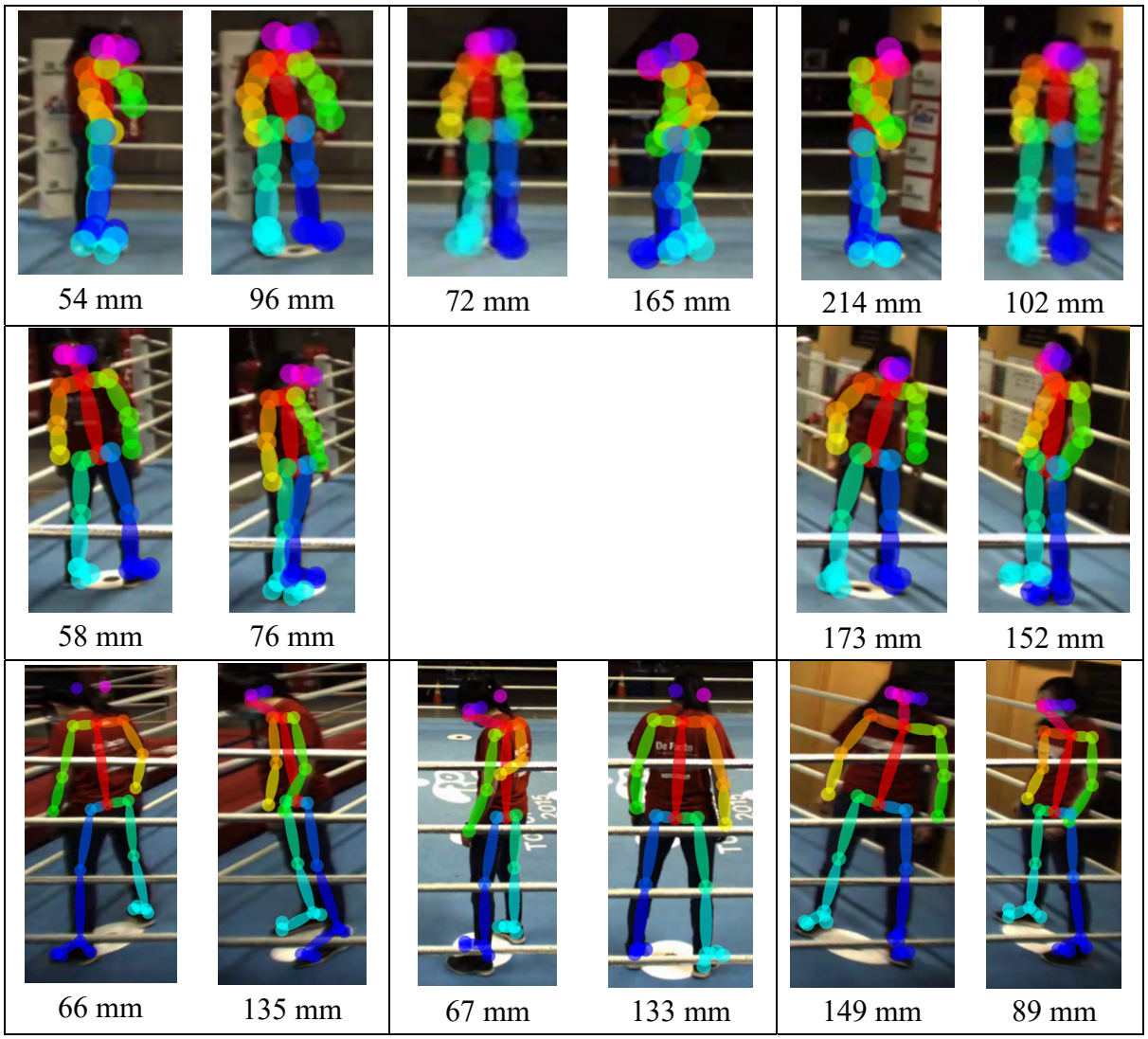


Figure 5.4 Erreur selon l'orientation du corps pour les repères près des cordes

5.2 Résultats qualitatifs

Cette section présente les résultats qualitatifs associés au classificateur et à l'algorithme de suivi des boxeurs. Elle comporte ainsi des exemples visuels de l'application de ces deux éléments pour l'analyse des performances.

5.2.1 Shadow boxing

La figure 5.5 illustre un exemple d'analyse pouvant être réalisée sur une vidéo de *shadow boxing*. La prédiction instantanée, située dans le coin supérieur gauche, indique l'action performée sur l'image actuelle. Le compteur, situé dans le coin supérieur droit, indique de manière incrémentale combien de coups de poing sont effectués. Finalement, le temps total passé en position ouverte ou en garde est inscrit dans le coin inférieur droit.

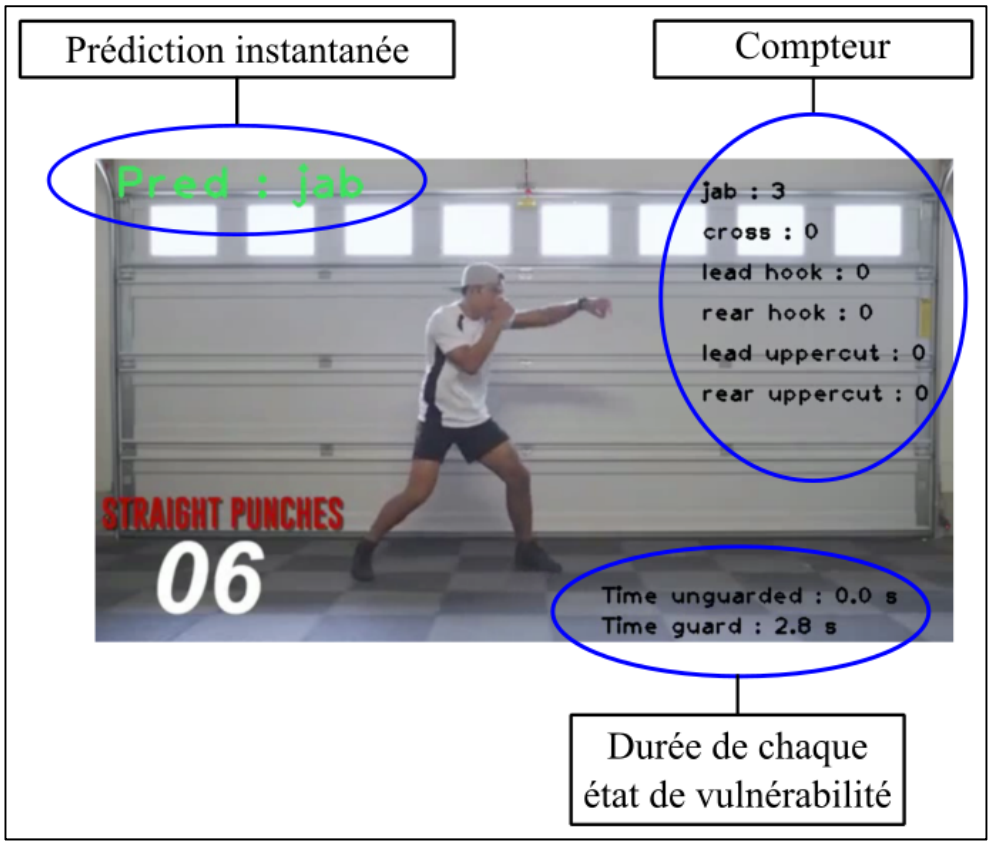


Figure 5.5 Exemple de résultat sur une vidéo de *shadow boxing*

5.2.2 Cartes de déplacement

L'algorithme de suivi des déplacements des boxeurs permet de produire les tracés de la position de chacun des athlètes durant le combat (Figure 5.6).

Figure 5.6 Déplacement total des boxeurs

Les cartes de chaleurs de la position des boxeurs peuvent de plus être générées (Figure 5.7 et Figure 5.8). En plus de permettre de visualiser la position des athlètes sur l'aire de combat, elles renseignent sur le temps passé à certains endroits spécifiques.

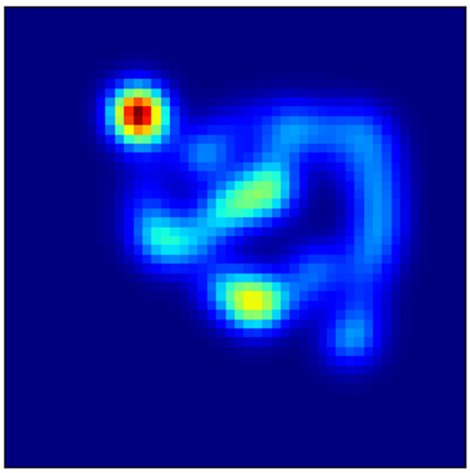


Figure 5.7 Carte de chaleur des déplacements – Boxeur 0

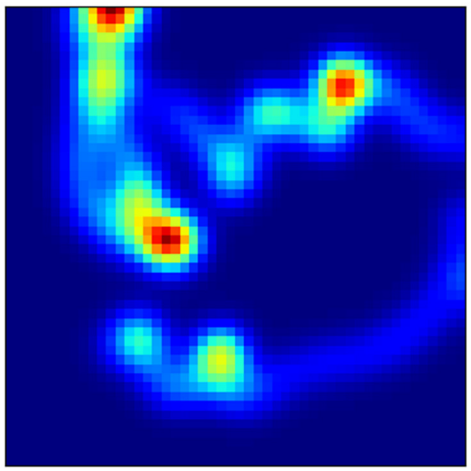
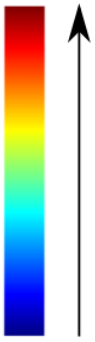


Figure 5.8 Carte de chaleur des déplacements – Boxeur 1



CHAPITRE 6

DISCUSSION

6.1 Comparaison avec la littérature

La portion de reconnaissance d'actions de la méthode d'extraction automatique des performances en boxe développée dans le cadre de cette maîtrise permet ainsi d'obtenir une exactitude de classification moyenne par action de 77 %. Cela est inférieur à celles d'études précédentes ayant cherché à classifier automatiquement les techniques offensives en boxe, qui se situent plutôt entre 86,7 % (Soekarjo *et al*., 2019) et 98 % (Worsey *et al*., 2020).

La première utilise les données issues de deux types de source, soit des images et des données de profondeur. Cette combinaison pourrait expliquer ces très bons résultats de classification. Quant aux deux autres, elles utilisent uniquement les données cinématiques provenant d'unités inertielles. Par contre, Worsey *et al*. (2020) entraînent et évaluent leur modèle avec des données provenant d'un seul même participant, ce qui aide fortement à obtenir une exactitude élevée. Quant à Khashanshin (2021), il s'agit de l'étude ayant le plus grand nombre d'exemples collectés, avec un total de 360 000, répartis hypothétiquement en 120 000 exemples pour chaque coup d'intérêt (direct, crochet et uppercut). En comparaison, le réseau LSTM bâti dans le cadre de cette maîtrise ne dispose que d'environ 1 768 exemples au total, pour un peu plus de 100 exemples de chaque technique offensive. Dans le domaine de l'apprentissage machine, le nombre de données est bien souvent le moteur des performances du réseau, et il est donc logique que le MLP de Khasanshin ait de très bonnes performances de classification.

L'exactitude de classification de la méthode développée dans le cadre de ce projet est donc inférieure à celles des méthodes usant de plusieurs sources de données, ou encore de capteurs inertiels. Toutefois, l'hypothèse que les performances de classification soient similaires à

celles de Khoury & Liu (2009) et Soekarjo *et al*. (2019) aurait pu être émise. En effet, ces deux études ayant utilisé les données tridimensionnelles issues de systèmes de capture de mouvement, ce qui est similaire aux poses extraites par un estimateur de pose 3D. Cela n'est cependant pas ce qui est observé, car l'exactitude obtenue de 77 % est près de 10 points de pourcentage inférieure à ces deux études (respectivement 87,71 % et 86,7 %). Les systèmes de capture de mouvement ayant une précision de l'ordre du micron tandis que les estimateurs de pose en ont plutôt une de l'ordre du centimètre, il est possible que cet écart explique celui des performances de classification. Néanmoins, il semble que l'utilisation des coordonnées 3D des parties du corps, qu'ils proviennent d'un système de capture de mouvement ou d'un estimateur de pose, ne permet pas d'atteindre des performances aussi élevées que d'autres méthodes utilisant des images vues de haut combinées à des données de profondeur, ou que des méthodes exploitant des capteurs inertiels. Puis, l'ensemble de données constitué dans le cadre de ce projet n'a pas la nature balancée de ceux des méthodes présentées dans la revue de la littérature. En effet, celles-ci ont pu contrôler le nombre de participants recrutés pour les séances d'acquisition de données, de même que le nombre et le type des techniques offensives effectuées par ceux-ci. Toutes les précédentes études ont un nombre égal de chaque technique offensive. Dans le cadre de ce projet, des vidéos YouTube contenant diverses personnes effectuant un nombre inégal de différents coups de poing ont plutôt été utilisées. Bien que celles-ci auraient pu être découpées de façon à obtenir un nombre égal de techniques offensives, la quantité de données a été priorisée sur le balancement de celles-ci. Le tableau 3.3 montre le nombre d'exemples moyen de chaque classe utilisés pour l'entraînement du classificateur. D'une part, il est possible de s'apercevoir que ces nombres sont inégaux à travers les classes. Il est également possible d'observer que le crochet arrière, l'uppercut avant et l'uppercut arrière sont des classes sous-représentées autant en entraînement qu'en évaluation. Cela se reflète dans les performances du réseau LSTM, puisqu'il s'agit des classes où ce dernier présente les plus basses métriques de classification. Le débalancement des classes est donc une autre piste d'explication des performances plus basses de notre méthode.

Une seconde différence de notre ensemble de données avec ceux des méthodes de la littérature est que ces dernières sont constituées d'actions individuelles déjà segmentées. Ce projet utilise plutôt des extraits de vidéos pouvant contenir plusieurs actions de différents types. Les méthodes de la littérature donnent donc une classe à une action, tandis que le LSTM entraîné dans le cadre de cette maîtrise peut détecter plusieurs actions lors d'un seul mouvement. Bien que les règles de filtrage des prédictions permettent d'éliminer une majorité de ces cas, le modèle peut continuer de détecter deux techniques offensives lorsqu'il n'y en a qu'une seule. Ce cas de figure est impossible à obtenir lorsque le classificateur est entraîné et évalué sur des actions déjà segmentées. Ainsi, le LSTM de ce projet est sujet à davantage de faux positifs, ce qui réduit la précision du modèle, et donc ses performances de classification.

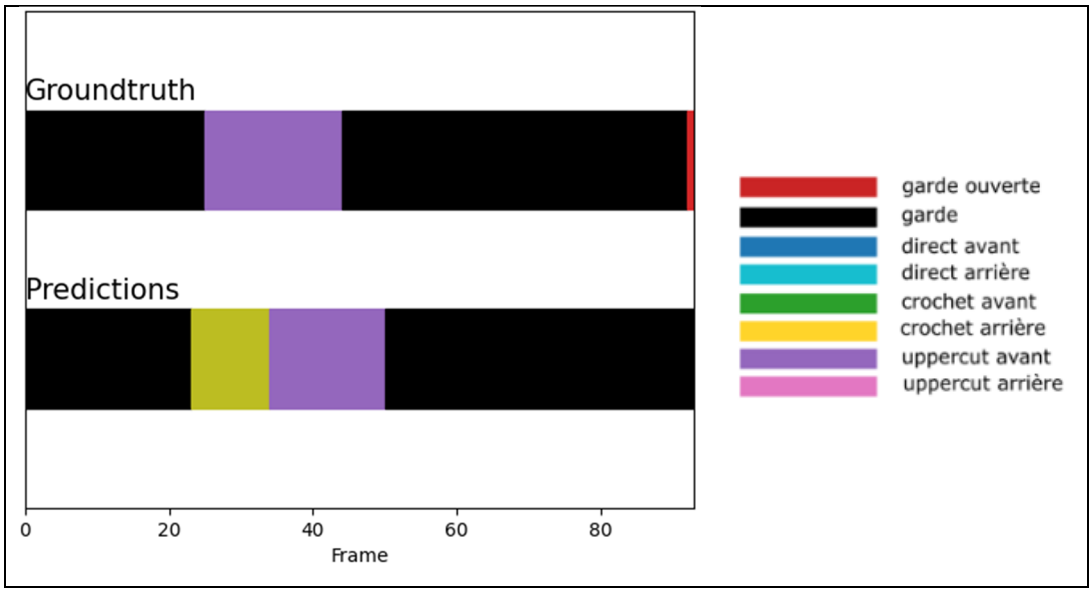


Figure 6.1 Détection de deux actions lors d'un seul mouvement

Bien que l'exactitude de classification du classificateur développé dans le cadre de cette maîtrise soit inférieure à celles des méthodes de classification des techniques de boxe de la littérature, la présente méthode offre l'avantage de ne nécessiter que des vidéos enregistrées par une simple caméra. Les méthodes plus performantes nécessitent de l'équipement spécialisé comme un système de capture de mouvement, une caméra à temps de vol ou des

capteurs inertiels. De plus, le besoin de logiciels spécialisés, de caméras infrarouges et de munir les athlètes de marqueurs est évité.

Un second avantage de la méthode faisant l'objet de ce mémoire est de pouvoir effectuer la reconnaissance d'actions sur des vidéos continues, ce qui est plus utile pour l'analyse de réels combats. Les études présentées dans à la section 1.1.3 de la revue de la littérature entraînent et évaluent leur modèle avec des extraits contenant une seule action à la fois. Cela implique donc que les coups de poing soient préalablement segmentés avant d'être classifiés, ce qui est effectué manuellement par les auteurs (Khoury & Liu, 2009; Kasiri-Bidhendi *et al*., 2015) ou automatiquement par seuillage de l'accélération mesurée par les capteurs inertiels (Worsey *et al*., 2019; Soekarjo *et al*., 2020; Khasanshin, 2021). Comme l'objectif est de ne pas devoir visionner et annoter manuellement les vidéos de combat pour en extraire le nombre et le type de coups de poing livrés dans un combat, en plus de ne pas obliger les athlètes à porter de l'équipement pendant leurs duels, ces méthodes ne sont pas pratiques pour l'analyse de réels combats.

6.2 Erreurs les plus fréquentes du classificateur

Les cas problématiques survenant le plus fréquemment sont les détections erronées de crochets arrière et d'uppercuts avant. Respectivement 22 % et 34 % du temps, ces techniques offensives sont détectées alors que le boxeur se trouve réellement en position de garde. La figure 6.2 et la figure 6.3 présentent deux moments où le coup de poing peut en effet ressembler à la position de garde, et donc mener à une classification erronée.

Figure 6.2 Crochet arrière classifié comme la position de garde

Figure 6.3 Uppercut avant classifié comme la position de garde

Le classificateur semble également avoir une certaine difficulté à faire la distinction entre les directs arrière et les crochets arrière. 15 % des directs arrière prédits sont en fait des crochets arrière, tandis qu'à l'inverse, 11 % des crochets arrière prédits sont réellement des directs arrière. La figure 6.4 et la figure 6.5 montrent des exemples où l'une de ces techniques offensives est classifiée comme l'autre. Bien que les poses adjacentes soient représentatives de la posture réelle du boxeur, sur ces images particulières, le poignet arrière est estimé un peu plus haut qu'il ne le devrait. Cela transforme donc l'allure du mouvement, ce qui peut expliquer la confusion entre ces deux actions déjà similaires dans leur initiation et leur retour vers la garde. Il est possible qu'un estimateur de pose 3D plus exact permette d'obtenir de meilleures performances de la part du classificateur.

Figure 6.4 Crochet arrière classifié comme un direct arrière

Figure 6.5 Direct arrière classifié comme un crochet arrière

Les erreurs de classification peuvent également survenir à cause des règles de filtrage des prédictions. Bien que celles-ci aient été choisies de manière à améliorer au mieux les performances de classification, il arrive qu'elles nuisent plutôt aux résultats. La figure 6.6 illustre un exemple où à trois reprises, le classificateur avait adéquatement détecté une partie du crochet arrière, mais où les règles de filtrage ont plutôt favorisé la classe « direct arrière ».

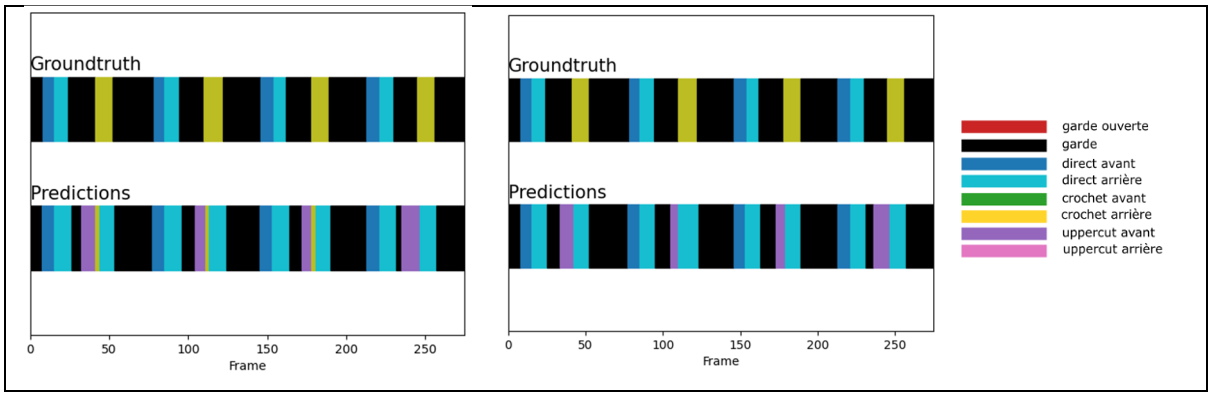


Figure 6.6 Prédictions avant (à gauche) et après (à droite) filtrage

La figure 6.7 illustre un exemple où les règles de filtrage retirent un faux positif, soit l'uppercut avant. Toutefois, elles enlèvent également la détection adéquate de l'uppercut arrière.

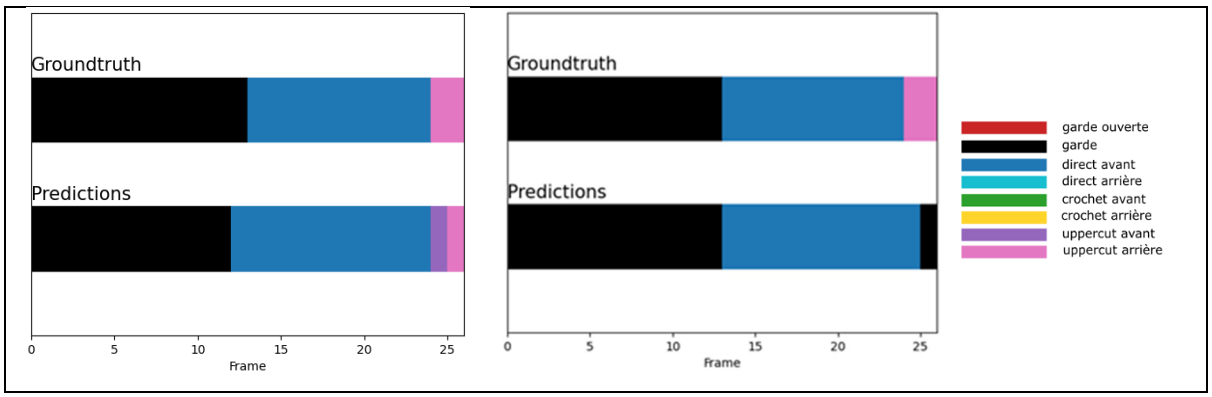


Figure 6.7 Uppercut arrière retiré après filtrage (à droite)

6.3 Utilité des métriques extraites pour l'analyse des performances en boxe

Tel que mentionné dans les études d'Ashker (2011), de Davis *et al*. (2015) et de Thompson & Lamb (2016), un nombre important de coups lancés dans un combat est corrélé avec un verdict de victoire. En effet, plus nombreuses sont les techniques offensives, plus grandes sont les chances d'en effectuer une qui atteint l'adversaire. La méthode développée lors de cette maîtrise permet d'obtenir facilement et rapidement ce nombre.

De plus, en combinant l'information du nombre et du type de coups lancés avec les déplacements du boxeur sur l'aire de combat, il est alors possible d'identifier son style de combat. Posséder cette information sur les athlètes adverses permet ainsi de dresser un portrait de ses tendances et préférences pour ensuite élaborer une stratégie appropriée.

6.4 Limitations et recommandations

Les performances de la méthode développée dans le cadre de cette maîtrise sont affectées par quelques limitations méthodologiques. Celles-ci sont présentées dans cette section, et des recommandations à leur sujet sont émises.

6.4.1 Ensemble de données

La section 6.1 a introduit trois limitations liées à l'ensemble de données spécifiques à la boxe constitué au cours de cette maîtrise, soit :

1) La taille de l'ensemble de données;
2) Le débalancement des classes;
3) La nature continue des extraits vidéos.

Aucun ensemble de données constitué de poses tridimensionnelles et dédié à la boxe n'étant accessible publiquement, cette présente maîtrise a nécessité la construction d'une telle base de données. Le projet étant restreint dans le temps, un nombre limité de vidéos de boxe a pu

être traité. Ainsi, l'ensemble de données utilisé pour entraîner le classificateur possède entre 100 et 200 exemples de chaque type de coup de poing. À titre comparatif, l'ensemble NTU RGB+D, qui est une base de données de référence pour la reconnaissance d'actions à partir de poses, contient 57 600 exemples, qui, hypothétiquement répartis également entre 60 classes d'action, représentent 948 exemples par action. Bien que l'ensemble de données spécifiques à la boxe ne contienne que huit actions par rapport à 60, les différentes techniques offensives ont plusieurs points communs tels que la vitesse du mouvement et l'implication majoritaire des membres supérieurs dans l'exécution de l'action. Khasanshin (2021) réussit d'ailleurs à classifier les directs, les crochets et les uppercuts avec une exactitude de 95,33 % grâce à un ensemble de données de 360 000 coups. Il est donc légitime de penser qu'augmenter la taille de la base de données mise sur pied durant cette maîtrise serait bénéfique pour les performances du classificateur. Ainsi, davantage de données, d'autant plus en contexte d'apprentissage machine, ne seraient que profitables pour les performances du classificateur.

En plus d'augmenter les données, il serait également pertinent de le faire de manière à ce qu'il y ait un nombre égal d'exemples pour chaque type d'actions d'intérêt. D'une part, cela permettrait une comparaison plus juste avec les méthodes de la littérature, qui possèdent toutes des ensembles de données balancés. D'autre part, il semble que les techniques offensives sous-représentées dans la base de données de ce présent projet sont celles les moins bien classifiées par le réseau de neurones. Bien que des poids aient été utilisés en entraînement pour compenser ce débalancement, cela ne semble pas suffisant pour obtenir des performances de classification homogènes.

Finalement, il faut tenir compte du fait que le réseau LSTM est entraîné et évalué avec des poses 3D provenant d'extraits vidéo continus pouvant contenir à la fois différentes actions et plusieurs exemples de celles-ci. Les précédentes méthodes observées dans la littérature emploient plutôt des extraits d'une action individuelle déjà segmentée. Le LSTM est donc sujet à générer davantage de faux positifs. Toutefois, étant donné que lors de l'analyse de

réels combats, les vidéos ne seront pas préalablement découpées en actions, puisque cela est en opposition directe avec l'objectif principal de ce projet, soit l'automatisation du processus d'analyse. Ainsi, l'utilisation de vidéos continues est plus représentative de l'application réelle de la méthode.

6.4.2 Étiquetage subjectif

L'étiquetage des données a été effectué par l'auteure de ce mémoire, n'ayant aucune formation préalable en ce qui a trait à la boxe. Les ouvrages de Werner & Lachica (2000), de Hatmaker & Werner (2004), de même que des tutoriels de boxe sur YouTube (dont sont issus la plupart des extraits retenus pour composer l'ensemble de données), ont été consultés afin d'assurer un étiquetage adéquat. Bien que les actions d'intérêt dans ce projet soient bien distinctes, il serait idéal de faire appel à des experts tels que des athlètes professionnels, des entraîneurs ou encore des juges pour valider l'étiquetage des poses tridimensionnelles.

6.4.3 Erreurs de l'estimateur de pose 3D

Le module de reconnaissance d'actions étant entraîné sur des données tridimensionnelles extraites par un estimateur de pose, les résultats de la classification sont dépendants de la performance de ce dernier. Utiliser un estimateur de pose 3D ayant la meilleure exactitude possible permet d'assurer une plus grande adéquation entre les poses et le mouvement d'origine. VIBE et PARE ont des performances compétitives avec les estimateurs de pose de l'état des connaissances, et les poses produites ont été inspectées visuellement pour s'assurer de leur qualité. Il est toutefois à noter que l'estimation de pose 3D n'atteint actuellement pas un niveau de précision équivalent à un système de capture de mouvement.

6.4.4 Algorithme de suivi des déplacements des boxeurs

L'algorithme de suivi des déplacements des boxeurs développé dans le cadre de cette maîtrise ne fonctionne que pour des vidéos ayant un point de vue fixe contenant l'entièreté de l'aire de combat. Les équipes d'excellence ayant pour habitude d'enregistrer leurs combats et

même leurs entraînements pour effectuer des analyses vidéo *a posteriori*, ces exigences peuvent être satisfaites plutôt facilement. Toutefois, il ne serait pas possible d'utiliser l'algorithme pour faire le suivi des boxeurs sur des vidéos de combats diffusés sur des chaînes sportives. Celles-ci ont en effet différents points de vue n'incluant pas toujours la totalité de l'aire de combat afin d'offrir une meilleure perspective de l'action pour les spectateurs. Dans ce cas particulier, il serait préférable d'utiliser les paramètres de caméra, qui doivent alors être connus ou estimés par des méthodes d'apprentissage machine (*e.g.* Bogdan *et al*., 2018), afin de positionner les boxeurs dans un système de coordonnées global.

6.5 Travaux futurs

Tout comme Khoury & Liu (2009), Kisiri-Bidhendi *et al*. (2015), Soejarko *et al*. (2019), Worsey *et al*. (2020) et Khasanshin (2021), la méthode d'extraction automatique de métriques de performances en boxe développée au cours de cette maîtrise ne permet la reconnaissance d'actions que sur des vidéos de *shadow boxing*, sur lesquelles n'apparaît qu'une seule personne. Dans des travaux futurs, la méthode sera adaptée et améliorée afin de pouvoir réaliser la classification d'actions sur des vidéos de combat entre deux boxeurs. Cela amènera de nouveaux défis comme l'occlusion entre plusieurs personnes et une plus grande diversité d'amplitude de mouvements selon que les coups atteignent ou non leur cible. L'évaluation du contact entre le boxeur donnant le coup et la cible fait également partie des défis à prévoir. Cette adaptation permettra en outre le calcul d'une plus grande variété de métriques de performance en boxe. En effet, en considérant l'interaction entre deux boxeurs, il est alors possible de mesurer l'efficience des coups lancés par l'athlète en déterminant si ces derniers atteignent leur cible, ou s'ils sont bloqués ou esquivés. Du même coup, l'efficience des techniques défensives pourra également être évaluée.

Une solution possible pour gérer l'occlusion est d'entraîner l'estimateur de pose 3D avec des données issues de plusieurs caméras ayant chacune un point de vue différent (Zhang *et al*., 2020). Le réentraînement du classificateur avec des poses 3D provenant de vidéos de combat entre deux boxeurs permettra également de couvrir la plus grande variété d'amplitude des

mouvements. Finalement, comme ces poses possèdent leur propre système de coordonnées, une partie du travail à faire sera d'utiliser les paramètres de caméra afin de les positionner précisément dans un référentiel commun. Puis, l'utilisation du modèle SMPL permettra de calculer l'interpénétration des parties du corps des boxeurs afin d'établir s'il y a contact ou non.

CONCLUSION

L'objectif de ce projet était de développer une méthode d'extraction automatique de métriques de performance en boxe. Des estimateurs de pose 3D ont donc été utilisés afin de constituer une base de données de poses tridimensionnelles destinées à l'entraînement d'un classificateur apte à distinguer la garde ouverte, la garde, le direct avant, le direct arrière, le crochet avant, le crochet arrière, l'uppercut avant et l'uppercut arrière avec une exactitude moyenne pondérée de 77 %. De plus, un algorithme permet de faire le suivi des boxeurs, à 12,2 cm près, afin de quantifier leur occupation de l'aire de combat. Contrairement aux méthodes retrouvées dans la littérature au moment de la rédaction de ce mémoire nécessitant systèmes de capture de mouvement, caméras à temps de vol ou capteurs inertiels, seules sont requises des images monoculaires enregistrées par une simple caméra. Cela est beaucoup moins coûteux que des systèmes de capture de mouvements, et évite le besoin d'équiper les boxeurs de marqueurs réfléchissants ou de capteurs. Il n'est pas nécessaire d'installer de matériel additionnel autour de l'aire de combat, ce qui permet d'appliquer aisément une telle méthode dans divers contextes, tant en combat à domicile qu'en compétition à l'externe.

Le projet ayant dû être mené à l'intérieur d'une période de deux ans, l'ampleur de la base de données ayant pu être constituée dans ce laps de temps représente une piste évidente d'amélioration, surtout lorsque comparée à d'autres ensembles de données dédiés à la reconnaissance d'actions. De plus, la suite logique de ce projet est d'adapter le module de reconnaissance d'actions à des combats entre deux athlètes. Des défis supplémentaires comme l'occlusion entre des personnes et la détermination du contact entre elles sont à prévoir. Toutefois, davantage de métriques de performance comme l'efficience des coups peuvent ainsi être extraites.

ANNEXE I

AUTRES TECHNIQUES OFFENSIVES

Le coup de poing descendant consiste en un mouvement semi-circulaire vertical. Le poing dépasse la tête du boxeur exécutant la technique offensive, puis redescend vers la mâchoire de l'adversaire. Ce coup permet de passer au-dessus de la garde de l'opposant (Figure-A I-1).

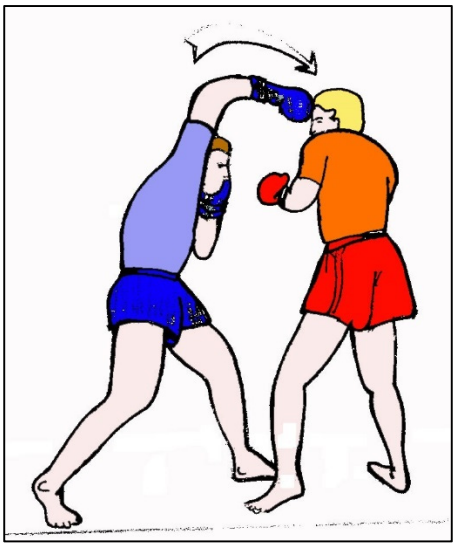


Figure-A I- 1 Coup de poing descendant
Tirée de Delmas (1999)

Le coup de poing Superman est une technique offensive effectuée dans les sports de combat où les coups de pieds sont permis, comme le *kickboxing* ou les arts martiaux mixtes. Le genou du côté de la main donnant le coup de poing est soulevé pour feinter un coup de pied, puis est ramené vers l'arrière dans un saut pendant qu'un direct est exécuté (Figure-A I- 2). La feinte de coup de pied vise à abaisser la garde averse afin de créer une ouverture pour le coup de poing. Cette technique offensive n'est généralement pas utilisée en boxe, puisqu'il n'est pas utile de feinter des coups de pied étant donné que ceux-ci sont interdits dans ce sport.

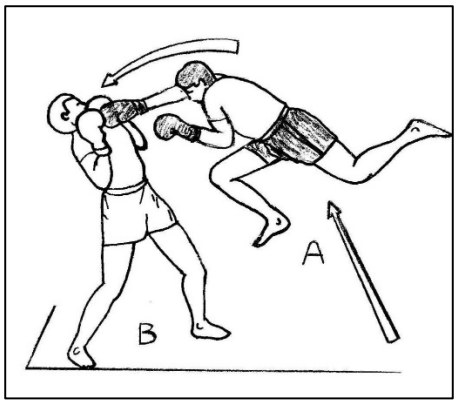


Figure-A I- 2 Coup de poing Superman
Tirée de Delmas (1999)

Le coup de poing retourné est donné avec le revers du poing en dépliant le coude (Figure-A I- 3). Il est à noter que cette technique offensive n'est pas permise dans les combats de boxe.

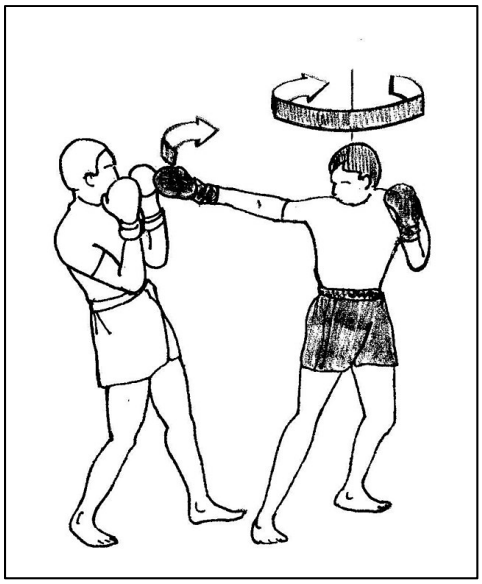


Figure-A I- 3 Coup de poing retourné
Tirée de Delmas (1999)

ANNEXE II

CONTENU DES ENSEMBLES DE DONNÉES DE RÉFÉRENCE

Tableau-A II- 1 Actions contenues dans l'ensemble de données UCF50

Interaction humain-objet	Cerceau	Balles de jonglerie	Corde à sauter	Mélangeur
	Nunchaku	Lancer de la pizza	Planche à roulettes	Jonglerie au soccer
	Yo-yo	-	-	-
Mouvement du corps	Saut en étoile	Fentes	Tractions	Pompes
	Escalade intérieure	Escalade cordes	Se balancer	Tai-chi
	Sauter sur une trampoline	Marcher avec un chien	-	-
Interaction humain-humain	Parade militaire	Pivot de salsa	-	-
Jouer d'un instrument de musique	Batterie	Guitare	Piano	Tabla
	Violon	-	-	-
Sports	Lancer au baseball	Lancer au basketball	Développé couché	Vélo
	Tir au billard	Brasse	Épaulé-jeté	Plonger
	Escrime	Coup au golf	Saut en hauteur	Course de chevaux
	Équitation	Lancer du javelot	Kayak	Saut à la perche
	Cheval d'arçon	Coup de poing	Skier	Motomarine
	Coup au tennis	Aviron	Lancer du disque	Frappe au volleyball

Tableau-A II- 2 Actions supplémentaires dans l'ensemble de données UCF101

Interaction humain-objet	Se maquiller	Se mettre du rouge à lèvres	Se sécher les cheveux	Se brosser les dents
	Couper des aliments dans la cuisine	Donner des coups de marteau	Tricoter	Laver le plancher avec une moppe
	Se raser la barbe	Écrire sur un clavier	Écrire sur un tableau	-
Mouvement du corps	Bébé qui rampe	Souffler des chandelles	Flexion sur jambes	Pompes en équilibre sur les mains
	Marcher sur les mains	Pompes au mur	-	-
Interaction humain-humain	Parade	Coupe de cheveux	Massage de tête	
Jouer d'un instrument de musique	Violoncelle	Daf	Dhol	Flûte
	Sitar	-	-	-
Sports	Tir à l'arc	Poutre d'équilibre	Dunk au basketball	Quilles
	Boxe sur sac de frappe	Boxe sur sac de frappe de vitesse	Plonger d'une falaise	Tir vers le guichet au cricket
	Tir au cricket	Tir de pénalité au hockey sur gazon	Gymnastique au sol	Attraper un frisbee
	Crawl	Lancer du marteau	Patinage artistique	Saut en longueur
	Barres parallèles	Rafting	Lancer du poids	Parachutisme
	Pénalité au soccer	Anneaux (gymnastique)	Combat de sumo	Surf
	Coup au tennis de table	Barres asymétriques	-	-

Tableau-A II- 3 Actions contenues dans l'ensemble de données NTU RGB+D

Actions quotidiennes	A1 : boire de l'eau	A2 : manger un repas	A3 : se brosser les dents	A4 : se brosser les cheveux
	A5 : lâcher un objet	A6 : ramasser un objet	A7 : lancer un objet	A8 : s'asseoir
	A9 : se lever	A10 : taper des mains	A11 : lire	A12 : écrire
	A13 : déchirer un papier	A14 : mettre une veste	A15 : retirer une veste	A16 : mettre un soulier
	A17 : retirer un soulier	A18 : mettre des lunettes	A19 : retirer des lunettes	A20 : mettre un chapeau ou une casquette
	A21 : retirer un chapeau ou une casquette	A22 : encourager	A23 : saluer de la main	A24 : donner un coup de pied à quelque chose
	A25 : fouiller dans sa poche	A26 : sautiller	A27 : sauter	A28 : téléphoner
	A29 : manipuler son téléphone ou sa tablette	A30 : écrire sur un clavier	A31 : pointer quelque chose	A32 : prendre un égoportrait
	A33 : regarder sa montre	A34 : se frotter les mains	A35 : hocher la tête de haut en bas	A36 : secouer la tête de gauche à droite
	A37 : s'essuyer le visage	A38 : salutation militaire	A39 : se mettre les mains paume contre paume	A40 : se croiser les bras
Actions médicales	A41 : éternuer/tousser	A42 : démarche ataxique	A43 : tomber au sol	A44 : avoir mal à la tête
	A45 : avoir mal à la poitrine	A46 : avoir mal au dos	A47 : avoir mal au cou	A48 : avoir la nausée/vomir
	A49 : s'éventer	-	-	-
Interactions avec soi ou une autre personne	A50 : donner un coup de poing ou une claque	A51 : donner un coup de pied	A52 : pousser	A53 : taper dans le dos
	A54 : pointer du doigt	A55 : donner un câlin	A56 : donner un objet	A57: toucher la poche de son pantalon
	A58 : serrer la main	A59 : marcher vers	A60 : s'éloigner en marchant	-

Tableau-A II- 4 Actions contenues dans l'ensemble de données NTU RGB+D

Actions quotidiennes	A61 : mettre des écouteurs	A62 : retirer des écouteurs	A63 : tir au basketball	A64 : faire rebondir un ballon au sol
	A65 : donner un coup de raquette	A66 : manipuler une balle de tennis de table	A67 : demander le silence	A68 : secouer les cheveux
	A69 : pouce en l'air	A70 : pouce vers le bas	A71 : signe du OK avec les doigts	A72 : faire un signe de victoire
	A73 : agrafer un livret	A74 : compter de l'argent	A75 : se couper les ongles	A76 : couper du papier
	A77 : claquer des doigts	A78 : ouvrir une bouteille	A79 : humer l'air/sentir	A80 : flexion des genoux
	A81 : lancer une pièce de monnaie	A82 : plier un papier	A83 : faire une boule de papier	A84 : jouer avec un cube magique
	A85 : appliquer de la crème sur le visage	A86 : appliquer de la crème sur les mains	A87 : mettre un sac	A88: retirer un sac
	A89 : mettre un objet dans un sac	A90 : retirer un objet d'un sac	A91 : ouvrir une boîte	A92: déplacer des objets lourds
	A93: secouer le poing	A94: lancer son chapeau ou sa casquette	A95: faire un signe d'abandon	A96: croiser les bras
	A97 : faire des cercles avec ses bras	A98 : croisement des bras (exercice)	A99 : courir sur place	A100 : talons-fesses
	A101 : touché de pied opposé	A102 : coup de pied latéral	-	-
Actions médicales	A103 : bâiller	A104 : s'étirer	A105 : se moucher	
Interactions avec soi ou une autre personne	A106: frapper avec un objet	A107: menacer au couteau	A108: faire tomber	A109: prendre un objet
	A110 : tirer avec un fusil	A111 : marcher sur le pied	A112: taper dans la main	A113: célébrer et boire
	A114 : transporter un objet	A115 : prendre une photo	A116 : suivre	A117 : chuchoter
	A118 : échanger des choses	A119 : soutenir quelqu'un	A120: jouer à roche-papier-ciseaux	

ANNEXE III

VIDÉOS YOUTUBE POUR L'ENSEMBLE DE DONNÉES

Acosta, G. [El Yuyu]. (2021, 19 juillet). *5 Ways To Set Up Rear Uppercut For Boxing | GIVEAWAY* [Vidéo en ligne]. Repéré à https://youtu.be/VQw2NBxXgWQ

[Boxing Academy LK]. (2020, 12 mai). *How to do the Rear Hook* [Vidéo en ligne]. Repéré à https://youtu.be/ig5lWbccpyc

Chiappetta, S. [Strength and Conditioning by Santo Chiappetta]. (2021, 25 février). *Level 1 Boxing Fundamentals - Rear Hook* [Vidéo en ligne]. Repéré à https://youtu.be/U-a0PGt-vo0

Fazen, S. [fightTIPS]. (2013, 13 août). *3 Best Boxing Punch Combos* [Vidéo en ligne]. Repéré à https://youtu.be/U-a0PGt-vo0

Fazen, S. [fightTIPS]. (2013, 5 août). *Learn How to Punch Like a Boxer* [Vidéo en ligne]. Repéré à https://youtu.be/MY9po0amDtg

Fazen, S. [fightTIPS]. (2013, 30 juillet). *The BEST Boxing Footwork Drill for Beginners* [Vidéo en ligne]. Repéré à https://youtu.be/3tTfUCuUFd0

Fazen, S. [fightTIPS]. (2016, 3 octobre). *5 Shadowboxing Drills for Footwork, Defense, & Cardio* [Vidéo en ligne]. Repéré à https://youtu.be/m8W0J49PSHY

Fazen, S. [fightTIPS]. (2020, 28 mai). *Solo Boxing Training for Core Strength & Footwork* [Vidéo en ligne]. Repéré à https://youtu.be/8AfFGuy4ODE

Fazen, S. [fightTIPS]. (2017, 22 juin). *3 Uppercut Variations: Long, Short, & The Bolo Punch* [Vidéo en ligne]. Repéré à https://youtu.be/0mYyUHGBHxY

[FRANK BOXING COACH]. (2021, 16 août). *How to: LEAD UPPERCUT (BOXING TUTORIALS) - FRANK BOXING COACH*. Repéré à https://youtu.be/C8XIKoNECfY

Gales, M. [Fight your way Fit]. (2015, 29 octobre). *How to Throw a Hook in Boxing (A step by step guide for Beginners)* [Vidéo en ligne]. Repéré à https://youtu.be/i8xrGCZdLJA

Gales, M. [Fight your way Fit]. (2016, 15 septembre). *How to Shadow Box for Beginners* [Vidéo en ligne]. Repéré à https://youtu.be/CEqIGeXZN_M

Gales, M. [Fight your way Fit]. (2019, 3 août). *The Heavy Bag for Beginners "A Complete Workout"* [Vidéo en ligne]. Repéré à https://youtu.be/eaGqOq0pNPY

Jeffries, T. [Tony Jeffries]. (2020, 20 juillet). *How to do Shadow Boxing for Beginners | Why Boxers Shadow Box* [Vidéo en ligne]. Repéré à https://youtu.be/i12A6G_LiJos

Jeffries, T. [Tony Jeffries]. (2021, 5 mai). *20 Minute Boxing Workout at Home | Boxercise* [Vidéo en ligne]. Repéré à https://youtu.be/XIHvfwYVy4A

Jeffries, T. [Tony Jeffries]. (2021, 23 septembre). *Perfect Punching Technique in Boxing #shorts* [Vidéo en ligne]. Repéré à https://www.youtube.com/shorts/sgPsg95Uoj0

Jeffries, T. [Tony Jeffries]. (2021, 6 novembre). **Secret* Tips For Throwing A Perfect Hook To The Body* [Vidéo en ligne]. Repéré à https://youtu.be/uXsOssBwLoA

Mehmedagic, N. [NeroMMA]. (2020, 28 juin). *How To Punch. Left Hook. Karate for MMA. Episode 28.* [Vidéo en ligne]. Repéré à https://youtu.be/wTtDboNc8pI

Mehmedagic, N. [NeroMMA]. (2020, 8 novembre). *Crossover Punch! 1.4 Million Views on IG. Viral 30 second Breakdown.* [Vidéo en ligne]. Repéré à https://www.youtube.com/shorts/TtMRSP0V8pE

Mehmedagic, N. [NeroMMA]. (2020, 22 novembre). *How To Throw A Rear Hook | Right Hook Tutorial* [Vidéo en ligne]. Repéré à https://youtu.be/l7WO09PPFzU

Mehmedagic, N. [NeroMMA]. (2020, 29 novembre). *How To Rear Uppercut | How to Throw a Rear Uppercut.* [Vidéo en ligne]. Repéré à https://youtu.be/l2pH85lxG4M

Mehmedagic, N. [NeroMMA]. (2020, 5 décembre). *How To Throw a Lead Uppercut | Lead Uppercut Tutorial.* [Vidéo en ligne]. Repéré à https://youtu.be/JLr852zUkE0

Mehmedagic, N. [NeroMMA]. (2021, 3 janvier). *Boxing combination | Uppercut to cross | 30 sec breakdown.* [Vidéo en ligne]. Repéré à https://www.youtube.com/shorts/f0z8XA7Opdw

Mehmedagic, N. [NeroMMA]. (2021, 17 mars). *Boxing sidestep. Feint combination.* [Vidéo en ligne]. Repéré à https://www.youtube.com/shorts/91-bI4BWE1w

Mehmedagic, N. [NeroMMA]. (2021, 20 juillet). *Fake jab – Hook combo. Boxing #shorts* [Vidéo en ligne]. Repéré à https://www.youtube.com/shorts/no0IcUCWnOc

[Metal Muscles]. (2021, 5 mars). *How To Throw A Rear Hook - Full guide and secret tips - [Learn Boxing]* [Vidéo en ligne]. Repéré à https://www.youtu.be/FSAVEVvVguE

Neal, E. [UNBROKEN FITNESS SOLUTIONS]. (2021, 14 mars). *Right Hook (Boxing Technique)* [Vidéo en ligne]. Repéré à https://youtu.be/e_fotambbfs

Scott, M. [McLeod Scott Boxing]. (2020, 5 décembre). *Rear check Hook || Soviet Breakdown | McLeod Scott Boxing* [Vidéo en ligne]. Repéré à https://youtu.be/CJS4zDeRS0I

Triana, M. [Ironboy Experience]. (2020, 25 mars). *15 MIN SHADOWBOXING WORKOUT SLIPPING PUNCHES* [Vidéo en ligne]. Repéré à https://youtu.be/SoXYNcIG4OE

Triana, M. [Ironboy Experience]. (2020, 26 mars). *15 MINUTE SHADOWBOXING WORKOUT BOB AND WEAVE* [Vidéo en ligne]. Repéré à https://youtu.be/ZyiBjLDKO04

Valtellini, J. [Bazooka Joe Valtellini]. (2017, 14 mars). *BKA - Episode #8 - The Rear Hook* [Vidéo en ligne]. Repéré à https://youtu.be/eXHwUu99sas

Valtellini, J. [Bazooka Joe Valtellini]. (2017, 28 mars). *BKA - Episode #10 - The Lead Uppercut* [Vidéo en ligne]. Repéré à https://youtu.be/FMsAtEUjWBo

Williams, D. [Dayne Williams]. (2018, 3 janvier). *Boxing Combo for Hardcore Right Hooks!!!* [Vidéo en ligne]. Repéré à https://youtu.be/NkFtQyzBdzE

Williams, T. [thomas williams]. (2019, 20 février). *left hook + right hook + boxing endurance !!* [Vidéo en ligne]. Repéré à https://youtu.be/a44KAYqmfhg

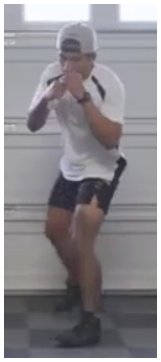
Yankello, T. [World Class Boxing Channel]. (2020, 19 avril). *Shadow Boxing Drill for Home | How To Defend and Counter Punches | Tom Yankello's Drill #3* [Vidéo en ligne]. Repéré à https://youtu.be/GP8f0mJlNFY

Yankello, T. [World Class Boxing Channel]. (2020, 23 avril). *Shadow Boxing Drill For Home | For the ELITE FIGHTER | Tom Yankello's Drill #4* [Vidéo en ligne]. Repéré à https://youtu.be/GP8f0mJlNFY

ANNEXE IV

EXEMPLES D'ÉTIQUETAGE

Tableau-A IV- 1 Exemples d'étiquetage des actions

Action	Exemple
garde ouverte	
garde	
direct avant	

Action	Exemple
direct arrière	
crochet avant	
crochet arrière	

Action	Exemple
uppercut avant	
uppercut arrière	

ANNEXE V

RECHERCHE EN GRILLE POUR LE RÉSEAU LSTM

x : pas d'apprentissage

Tableau-A V- 1 Comparaison des régularisations L1 et L2

Métrique	Perte (↓)	Exactitude (↑)	Score F1 macro (↑)
L1(0,0001)	2,45	0,75	0,58
L1(0,001)	**1,87**	**0,79**	**0,67**
L1(0,01)	x	x	x
L1(0,1)	x	x	x
L2(0,0001)	2,36	**0,79**	0,62
L2(0,001)	1,98	0,76	0,62
L2(0,01)	2,37	0,77	0,61
L2(0,1)	x	x	x

Tableau-A V- 2 Comparaison du taux d'apprentissage (lr)

Métrique	**lr = 0,001**	**lr = 0,01**	**lr = 0,1**
Perte (↓)	**1,81**	1,87	x
Exactitude (↑)	0,75	**0,79**	x
Score F1 macro (↑)	0,62	**0,67**	x

Tableau-A V- 3 Comparaison de la valeur de momentum

Métrique	**momentum = 0,9**	**momentum = 0,95**	**momentum = 0,97**
Perte (↓)	2,21	1,87	**1,17**
Exactitude (↑)	0,75	**0,79**	0,77
Score F1 macro (↑)	0,56	**0,67**	0,66

Tableau-A V- 4 Comparaison des optimiseurs SGD et Adam

Métrique	SGD	Adam		
	lr = 0,01, momentum = 0,95	lr = 0,0001	lr = 0,001	lr = 0,01
Perte (↓)	1,87	**1,64**	1,84	1,68
Exactitude (↑)	**0,79**	0,75	0,78	0,68
Score F1 macro (↑)	**0,67**	0,61	0,60	0,51

Tableau-A V- 5 Comparaison d'une couche LSTM vs deux couches LSTM

Métrique	1 couche LSTM	2 couches LSTM
Perte (↓)	**1,87**	3,01
Exactitude (↑)	**0,79**	0,76
Score F1 macro (↑)	**0,67**	0,55

Tableau-A V- 6 Comparaison du nombre d'unités de la couche LSTM

Métrique	LSTM 64	LSTM 128	LSTM 256
Perte (↓)	2,05	**1,87**	2,68
Exactitude (↑)	0,79	0,79	**0,81**
Score F1 macro (↑)	0,65	**0,67**	0,64

Tableau-A V- 7 Comparaison du nombre d'unités de la couche Dense

Métrique	Dense 32	Dense 64	Dense 128
Perte (↓)	2,04	1,87	**1,82**
Exactitude (↑)	0,76	**0,79**	**0,79**
Score F1 macro (↑)	0,60	**0,67**	0,62

Tableau-A V- 8 Comparaison de la taille d'un lot

Métrique	16	32	64
Perte (↓)	1,94	**1,87**	2,02
Exactitude (↑)	0,77	0,79	**0,80**
Score F1 macro (↑)	0,63	**0,67**	0,65

Tableau-A V- 9 Effet du nombre de parties du corps pour la classification

Métrique	15 parties du corps	25 parties du corps
Perte (↓)	**1,87**	2,15
Exactitude (↑)	**0,79**	0,76
Score F1 macro (↑)	**0,67**	0,59

ANNEXE VI

VALIDATION CROISÉE À 5 PLIS

Tableau-A VI- 1 Pli 1 de la validation croisée

Classe	Précision (%)	Rappel (%)	Score F1 (%)	Vrais positifs	Faux positifs	Faux négatifs	Nombre d'exemples
Garde ouverte	100	50	67	1	1	0	2
Garde	98	84	90	90	17	2	107
Direct avant	67	100	80	16	0	8	16
Direct arrière	73	100	84	8	0	3	8
Crochet avant	50	91	65	10	1	10	11
Crochet arrière	47	100	64	17	0	19	17
Uppercut avant	64	100	78	21	0	12	21
Uppercut arrière	94	55	70	16	13	1	29

Tableau-A VI- 2 Pli 2 de la validation croisée

Classe	Précision (%)	Rappel (%)	Score F1 (%)	Vrais positifs	Faux positifs	Faux négatifs	Nombre d'exemples
Garde ouverte	100	100	100	2	0	0	2
Garde	93	84	88	202	38	16	240
Direct avant	98	81	89	44	10	1	54
Direct arrière	80	87	83	59	9	15	68
Crochet avant	86	70	77	70	30	11	100
Crochet arrière	0	0	0	0	18	3	18
Uppercut avant	80	28	42	16	41	4	57
Uppercut arrière	90	54	68	27	23	3	50

Tableau-A VI- 3 Pli 3 de la validation croisée

Classe	Précision (%)	Rappel (%)	Score F1 (%)	Vrais positifs	Faux positifs	Faux négatifs	Nombre d'exemples
Garde ouverte	0	0	0	0	2	0	2
Garde	97	81	88	86	20	3	106
Direct avant	94	85	89	17	3	1	20
Direct arrière	100	100	100	11	0	0	11
Crochet avant	83	83	83	24	5	5	29
Crochet arrière	83	56	67	10	8	8	18
Uppercut avant	53	80	64	8	2	2	10
Uppercut arrière	46	100	63	11	0	13	11

Tableau-A VI- 4 Pli 4 de la validation croisée

Classe	Précision (%)	Rappel (%)	Score F1 (%)	Vrais positifs	Faux positifs	Faux négatifs	Nombre d'exemples
Garde ouverte	0	25	0	1	3	0	4
Garde	96	66	78	152	80	6	232
Direct avant	95	95	95	55	3	3	58
Direct arrière	76	97	86	71	2	22	73
Crochet avant	93	87	90	39	6	3	45
Crochet arrière	91	30	45	10	23	1	33
Uppercut avant	27	91	42	10	1	27	11
Uppercut arrière	55	100	71	11	0	9	11

Tableau-A VI- 5 Pli 5 de la validation croisée

Classe	Précision (%)	Rappel (%)	Score F1 (%)	Vrais positifs	Faux positifs	Faux négatifs	Nombre d'exemples
Garde ouverte	0	57	0	4	3	2	7
Garde	94	76	84	136	43	9	179
Direct avant	86	60	70	31	21	5	52
Direct arrière	37	37	37	7	12	12	19
Crochet avant	98	85	91	46	8	1	54
Crochet arrière	26	60	36	6	4	17	10
Uppercut avant	18	20	19	2	8	9	10
Uppercut arrière	59	63	61	22	13	15	35

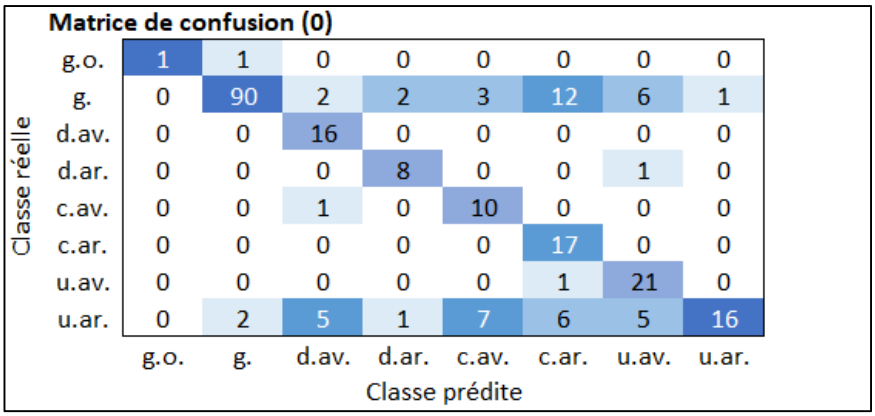


Figure-A VI- 1 Matrice de confusion 0

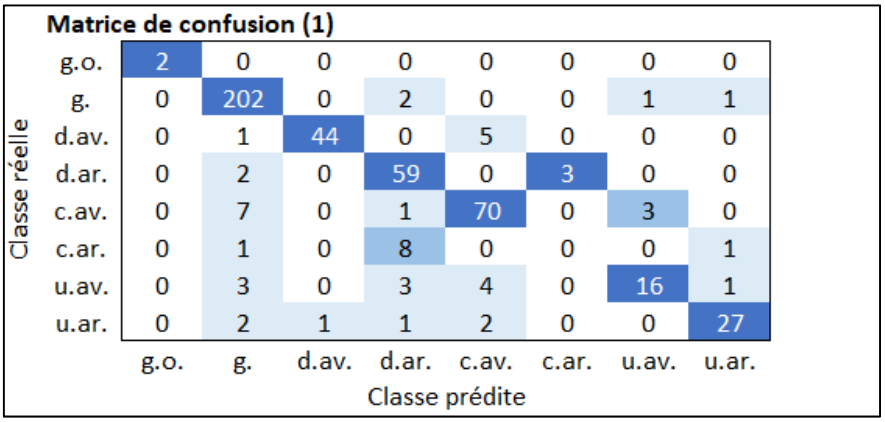


Figure-A VI- 2 Matrice de confusion 1

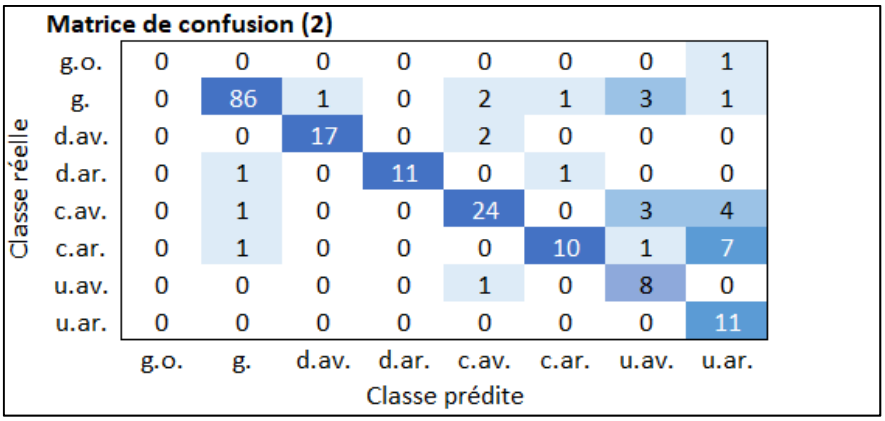


Figure-A VI- 3 Matrice de confusion 2

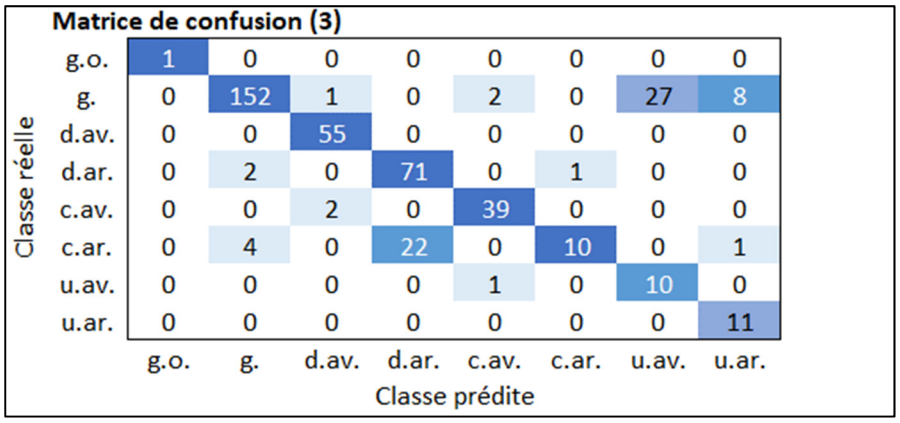


Figure-A VI- 4 Matrice de confusion 3

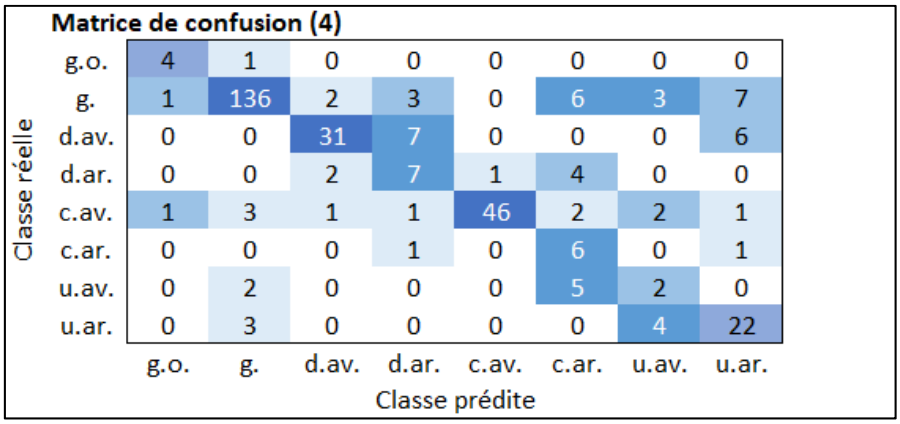


Figure-A VI- 5 Matrice de confusion 4

ANNEXE VII

VALIDATION DE L'ALGORITHME DE SUIVI

Tableau-A VII- 1 Résultats complets de la validation de l'algorithme de suivi

# Repère	Position réelle (x, y) (mm)	Positions algorithme (x, y) (mm)		Erreur moyenne (mm)	Écart-type (mm)
		Test #1	Test #2		
1	(500, 500)	(543, 468) (529, 408)	(519, 470) (634, 581)	86	54
2	(3048, 500)	(2996, 550) (3193, 421)	(2999, 492) (3091, 544)	87	53
3	(5596, 500)	(5383, 526) (5504, 457)	(5450, 577) (5445, 632)	171	51
4	(5596, 3048)	(5563, 2878) (5653, 2907)	(5585, 2876) (5492, 2879)	174	19
5	(5596, 5596)	(5635, 5453) (5636, 5517)	(5611, 5474) (5637, 5537)	108	34
6	(3048, 5596)	(3014, 5538) (2915, 5594)	(2927, 5467) (3017, 5469)	127	45
7	(500, 5596)	(449, 5554) (433, 5479)	(409, 5498) (333, 5552)	127	44
8	(500, 3048)	(467, 3001) (440, 3001)	(441, 3029) (485, 3027)	55	21
9	(2048, 3048)	(1933, 2994) (2057, 3019)	(2014, 2963) (1911, 2965)	102	56
10	(2048, 2048)	(1867, 2026) (1982, 1979)	(2045, 2207) (1942, 1904)	154	40
11	(3048, 2048)	(2916, 1819) (3019, 2020)	(2989, 1941) (2977, 1980)	131	95
12	(4048, 2048)	(3932, 1866) (4071, 1980)	(3921, 1865) (4032, 1942)	154	76
13	(4048, 3048)	(3959, 2945) (4041, 2975)	(3946, 2918) (4014, 3003)	108	52
14	(4048, 4048)	(4013, 3902) (3987, 4026)	(3960, 4065) (3877, 4105)	121	53
15	(3048, 4048)	(3075, 3782) (2971, 4072)	(2987, 4052) (3036, 3952)	126	95
16	(2048, 4048)	(2054, 3919) (1995, 3960)	(2032, 3939) (1922, 3941)	127	28
MOYENNE				122	57

ANNEXE VIII

COMPARAISON COMPLÈTE DES ORIENTATIONS DU CORPS

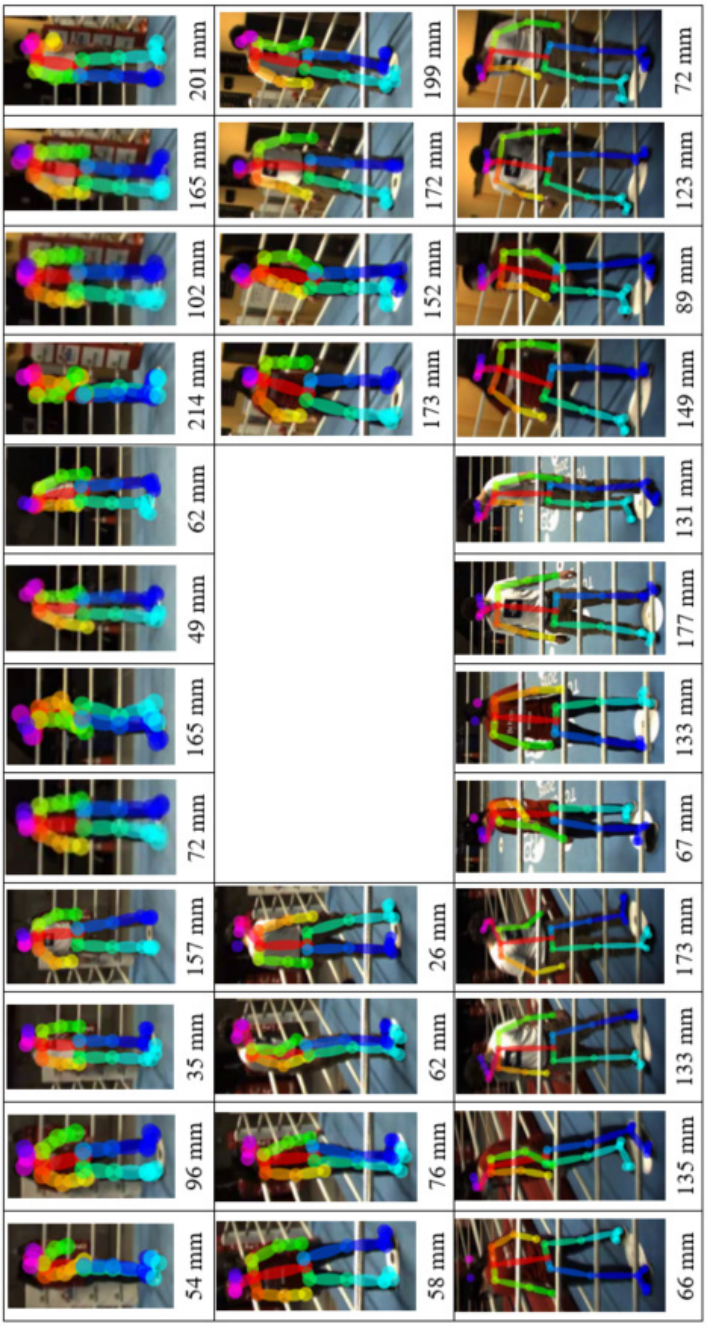


Figure-A VIII- 1 Validation, pourtour de l'aire de combat

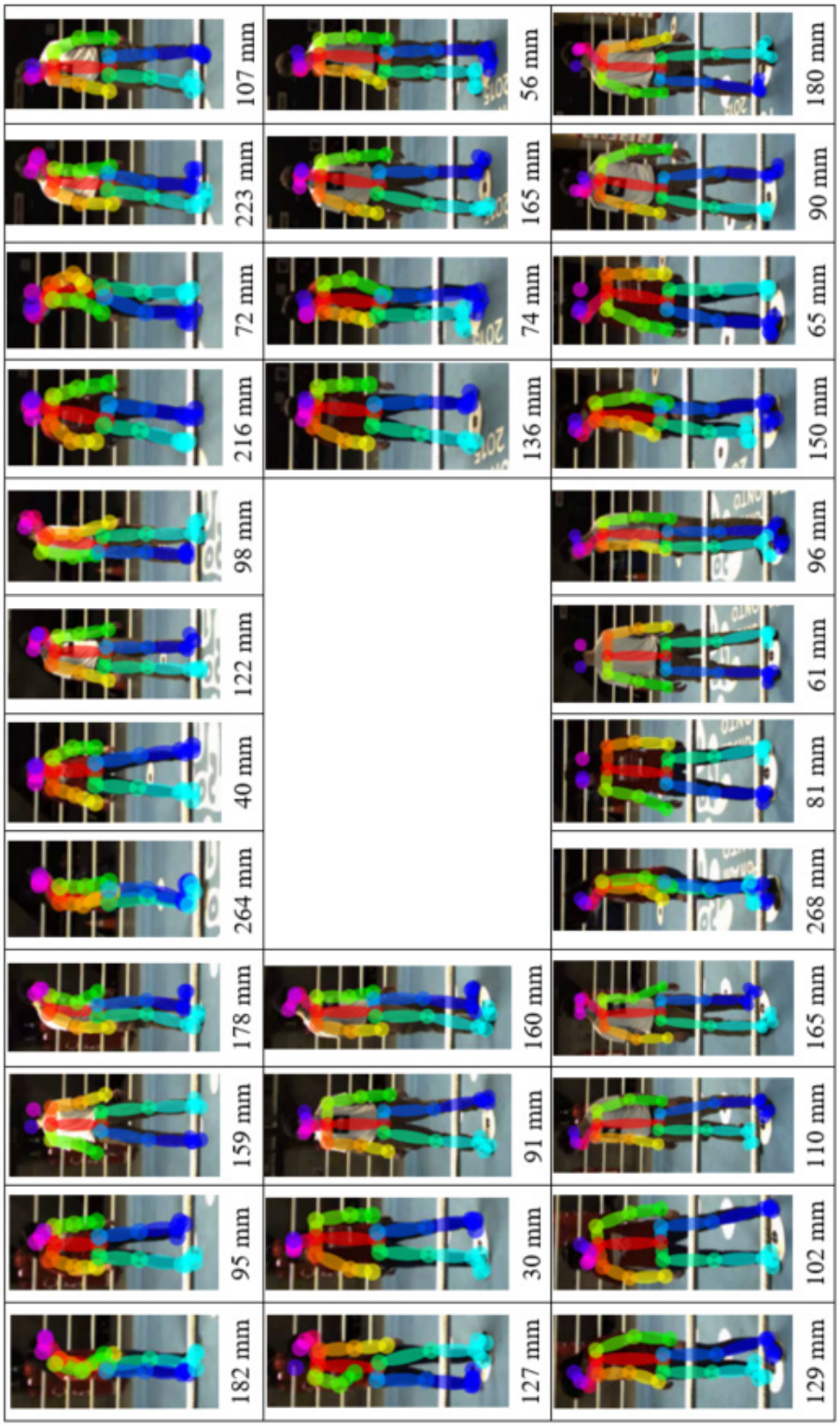


Figure-A VIII- 2 Validation, centre de l'aire de combat

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

Ashker, S. E. (2011). Technical and tactical aspects that differentiate winning and losing performances in boxing. *International Journal of Performance Analysis in Sport, 11*(2), 356-364. doi: 10.1080/24748668.2011.11868555

Bogdan, O., Eckstein, V., Rameau, F. & Bazin, J-C. (2018). DeepCalib: A Deep Learning Approach for Automatic Intrinsic Calibration of Wide Field-of-View Cameras. *Proceedings of the ACM SIGGRAPH European Conference on Visual Media Production*, *CVMP 2018, London, United Kingdom, December 13-14, 2018*, 6:1-6:10. doi: 10.1145/3278471.3278479

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., & Black, M. J. (2016). Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. *Proceedings of the European Conference on Computer Vision, ECCV 2016, Amsterdam, The Netherlands, October 8-16, 2016*, 561-578. doi: 10.1007/978-3-319-46454-1_34

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(1), 172-186. doi: 10.1109/TPAMI.2019.2929257

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., & Zisserman, A. (2018). *A Short Note about Kinetics-600*. Repéré à http://arxiv.org/abs/1808.01340

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., & Sun, J. (2018). Cascaded Pyramid Network for Multi-Person Pose Estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, Utah, June 18-23, 2018*, 7103-71212. doi: 10.1109/CVPR.2018.00742

Cheng, Y., Yang, B., Wang, B., Wending, Y., & Tan, R. (2019). Occlusion-Aware Networks for 3D Human Pose Estimation in Video. *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea*, *2019*, 723-732. doi: 10.1109/ICCV.2019.00081

Cheng, Y., Yang, B., Wang, B., & Tan, R. T. (2020). 3D Human Pose Estimation using Spatio-Temporal Networks with Explicit Occlusion Training. *Proceedings of the AAAI Conference on Artificial Intelligence, New York, United States of America, February 7-12, 2020*, 10631-10638. doi: https://doi.org/10.1609/aaai.v34i07.6689

Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., & Lu, H. (2020). Skeleton-Based Action Recognition With Shift Graph Convolutional Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, United States of America*, *June 14-19*, *2020*, 183-192. doi: 10.1109/CVPR42600.2020.00026

Davis, P., Benson, P. R., Pitty, J. D., Connorton, A. J., & Waldock, R. (2015). The Activity Profile of Elite Male Amateur Boxing. *International Journal of Sports Physiology and Performance, 10*(1), 53-57. doi: 10.1123/ijspp.2013-0474

Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhagen, R., & Van Gool, L. (2020). Large Scale Holistic Video Understanding. *Proceedings of the European Conference on Computer Vision, ECCV 2020,* [En ligne]*, August 23-28, 2020*, 593-610. doi: 10.1007/978-3-030-58558-7_35

Godwisdom. (2020, 3 septembre). SMPL-X [Billet de blogue]. Repéré à https://blog.csdn.net/u011681952/article/details/103768018?spm=1001.2014.3001.5501

Hatmaker, M. & Werner, D. (2004). Boxing Mastery: Advanced Technique, Tactics and Strategies from the Sweet Science. Tracks Publishing, California.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, United States of America, June 26th – July 1st, 2016*, 770-778. doi: 10.1109/CVPR.2016.90

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-term Memory. *Neural Computation, 9*(8), 1735-1780. doi: 10.1162/neco.1997.9.8.1735

Hu, W., Zhang, C., Zhan, F., Zhang, L., & Wong, T.-T. (2021). Conditional Directed Graph Convolution for 3D Human Pose Estimation. *Proceedings of the ACM International Conference on Multimedia, MM 2021, Chengdu, China, October 20-24, 2021*, 602-611. doi: 10.1145/3474085.3475219

International Boxing Association (2021). IBA Technical and Competition Rules.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2014). Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(7), 1325-1339. doi: 10.1109/TPAMI.2013.248

Kasiri-Bidhendi, S., Fookes, C., Morgan, S., Martin, D. T., & Sridharan, S. (2015). Combat sports analytics: Boxing punch classification using overhead depth imagery. *Proceedings of the IEEE International Conference on Image Processing, ICIP 2015, Quebec, Canada, 2015*, 4545-4549. doi: 10.1109/ICIP.2015.7351667

Kasiri, S., Fookes, C., Sridharan, S. & Morgan, S. (2017). Fine-grained action recognition of boxing punches from depth imagery. *Computer Vision and Image Understanding, 159*, 143-153. doi: 10.1016/j.cviu.2017.04.007

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S. & Zisserman, A. (2017). *The Kinetics Human Action Video Dataset*. Repéré à https://arxiv.org/abs/1705.06950

Khasanshin, I. (2021). Application of an Artificial Neural Network to Automate the Measurement of Kinematic Characteristics of Punches in Boxing. *Applied Sciences, 11*(3). doi: 10.3390/app11031223

Khoury, M., & Liu, H. (2009). Boxing Motions Classification through Combining Fuzzy Gaussian Inference with a Context-Aware Rule-Based System. *Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2019, Jeju, Korea, 2019*, 842-847. doi: 10.1109/FUZZY.2009.5277351

Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. *Proceedings of the International Conference on Learning Representation, ICLR 2017, Toulon, France, April 24-26, 2017*.

Kocabas, M., Athanasiou, N., & Black, M. J. (2020). VIBE: Video Inference for Human Body Pose and Shape Estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2020,* [En ligne]*, June 14-19, 2020*, 5253-5263. doi: 10.1109/CVPR42600.2020.00530

Kocabas, M., Huang, C.-H. P., Hilliges, O., & Black, M. J. (2021). PARE: Part Attention Regressor for 3D Human Body Estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2021,* [En ligne]*, October 11-17, 2021,* 11127-11137. doi: 10.1109/ICCV48922.2021.01094

Kolotouros, N., Pavlakos, G., Black, M. J., & Daniilidis, K. (2019). Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, October 27th – November 2nd, 2019*, 2252-2261. doi: 10.1109/ICCV.2019.00234

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal Convolutional Networks for Action Segmentation and Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, Hawaii, July 21-26, 2017*, 1003-1012. doi: 10.1109/CVPR.2017.113

Li, Y., Lu, Z., Xiong, X., & Huang, J. (2022). PERF-Net: Pose Empowered RGB-Flow Net. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WAVC 2022, Waikoloa, Hawaii, January 4-8, 2022*, 513-522. doi: 10.1109/WACV51458.2022.00087

Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.-Y., & Kot, A. C. (2020). NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(10), 2684-2701. doi: 10.1109/TPAMI.2019.2916873

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics, 34*(6), 1-16. doi: 10.1145/2816795.2818013

Martinez, J, Hossain, R., Romero, J. & Little, J. J. (2017). A Simple yet Effective Baseline for 3D Human Pose Estimation. *Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29*, 2640-2649. doi: 10.1109/ICCV.2017.288

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., & Theobalt, C. (2017). Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. *Proceedings of the International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, 506-516. doi: 10.1109/3dv.2017.00064

Moon, G., & Lee, K. M. (2020). I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. *Proceedings of the European Conference on Computer Vision, ECCV 2020,* [En ligne]*, August 23-28, 2020*, 752-768. doi: 10.1007/978-3-030-58571-6_44

Nibali, A., He, Z., Morgan, S., & Prendergast, L. (2019). 3D Human Pose Estimation with 2D Marginal Heatmaps. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WAVC 2019, Waikoloa, Hawaii, January 7-11, 2019*, 1477-1485. doi: 10.1109/WACV.2019.00162

Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K. (2017). Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, Hawaii, July 21-26, 2017*, 7025-7034. doi: 10.1109/CVPR.2017.139

Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). 3D human pose estimation in video with temporal convolutions and semi-supervised training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, United States of America, June 16-20, 2019*, 7753-7762. doi: 10.1109/CVPR.2019.00794

Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., & Schiele, B. (2016). DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, United States of America, June 26th – July 1st, 2016*, 4929-4937. doi: 10.1109/CVPR.2016.533

Qiu, Z., Yao, T., Ngo, C.-W., Tian, X., & Mei, T. (2019). Learning Spatio-Temporal Representation with Local and Global Diffusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, United States of America, June 16-20, 2019*, 12048-12057. doi: 10.1109/CVPR.2019.01233

Reddy, K. K., & Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine Vision and Applications, 24*(5), 971-981. doi: 10.1007/s00138-012-0450-4

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1987). Learning Internal Representations by Error Propagation. MIT Press. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, 318-362.

Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, United States of America, June 26th – July 1st, 2016*, 1010-1019. doi: 10.1109/CVPR.2016.115

Si, C., Chen, W., Wang, W., Wang, L., & Tan, T. (2019). An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, United States of America, June 16-20, 2019*, 1227-1236. doi: 10.1109/CVPR.2019.00132

Sigal, L., Balan, A. O., & Black, M. J. (2010). HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *International Journal of Computer Vision, 87*(1), 4-27. doi: 10.1007/s11263-009-0273-6

Soekarjo, K. M. W., Orth, D., Warmerdam, E., & van der Kamp, J. (2019). Automatic Classification of Strike Techniques Using Limb Trajectory Data. Dans U. Brefeld, J. Davis, J. Van Haaren, & A. Zimmermann (Éds), *Machine Learning and Data Mining for Sports Analytics* (Vol. 11330, pp. 131-141). Cham: Springer International Publishing. doi: 10.1007/978-3-030-17274-9_11

Soomro, K., Zamir, A. R. & Shah, M. (2012). *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*. Repéré à https://arxiv.org/abs/1212.0402

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep High-Resolution Representation Learning for Human Pose Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, United States of America, June 16-20, 2019*, 5693-5703. doi: 10.1109/CVPR.2019.00584

Thomson, E., & Lamb, K. (2016). The technical demands of amateur boxing: Effect of contest outcome, weight and ability. *International Journal of Performance Analysis in Sport, 16*(1), 203-215. doi: 10.1080/24748668.2016.11868881

Tripathi, A. (2021). What is the Main Difference between RNN and LSTM | NLP | RNN vs LSTM. Data Science Duniya. Repéré à https://ashutoshtripathi.com/2021/07/02/what-is-the-main-difference-between-rnn-and-lstm-nlp-rnn-vs-lstm/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the Conference on Neural Information Processing Systems, NIPS 2017, Long Beach, United States of America, December 4-9, 2017*, 6000-6010. doi: 10.5555/3295222.3295349

von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., & Pons-Moll, G. (2018). Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. Dans V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Éds), *Computer Vision – ECCV 2018* (Vol. 11214, pp. 614-631). Cham: Springer International Publishing. doi: 10.1007/978-3-030-01249-6_37

Werner, D. & Lachica, A. (2000). Fighting Fit: Boxing Workouts Techniques and Sparring. Tracks Publishing, California.

Worsey, M. T. O., Espinosa, H. G., Shepherd, J. B., & Thiel, D. V. (2020). An Evaluation of Wearable Inertial Sensor Configuration and Supervised Machine Learning Models for Automatic Punch Classification in Boxing. *IoT, 1*(2), 360-381. doi: 10.3390/iot1020021

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1). doi: 10.1609/aaai.v32i1.12328

Zhao, R., Wang, K., Su, H., & Ji, Q. (2019). Bayesian Graph Convolution LSTM for Skeleton Based Action Recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV 2019*, 6882-6892. doi: 10.1109/ICCV.2019.00698

Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N. (2017). View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. *Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2117-2126. doi: 10.1109/ICCV.2017.233

Zhang, J., Tu, Z., Yang, J., Chen, Y., & Yuan, J. (2022). MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, Louisiana, June 16-24, 2022*, 13232-13242.

Zhou, X., Sun, X., Zhang, W., Liang, S., & Wei, Y. (2016). Deep Kinematic Pose Regression. *Proceedings of the European Conference on Computer Vision, ECCV 2016, Amsterdam, The Netherlands, October 8-16, 2016*. doi: 10.1007/978-3-319-49409-8_17