Projet MIDL 2

CROS Héléna
HO Sylvie
LORY Solène
SALLET Eloïse

Projet de la double licence mathématiques et informatique
Janvier 2026

Table des matières

1 Introduction

L'analyse automatisée du mouvement humain représente aujourd'hui un enjeu majeur de la vision par ordinateur, avec des applications allant de la biomécanique au divertissement numérique. Dans le cadre de ce projet, notre objectif est d'étudier la capacité des modèles d'intelligence artificielle à extraire, comparer et prédire des séquences chorégraphiques complexes à partir de simples flux vidéos.

Nous avons porté notre choix sur la "Danse de l'Égyptien", une chorégraphie issue du jeu Just Dance. Ce choix se justifie par la nature hautement géométrique de cette danse. Caractérisée par des postures angulaires et des segments corporels souvent parallèles ou perpendiculaires, elle impose des contraintes stylistiques strictes. Ces particularités constituent un terrain d'expérimentation idéal pour évaluer la précision de nos algorithmes d'extraction de pose et de prédiction de trajectoires.

Notre démarche s'articule autour de trois axes principaux, débutant par l'extraction de données structurées à partir de vidéos de référence, suivie d'une analyse comparative entre une performance professionnelle et notre propre reproduction du mouvement effectuée par notre groupe de quatre danseurs, pour aboutir enfin à la génération prédictive de mouvements via des réseaux de neurones afin de simuler la suite logique d'une chorégraphie commencée.

2 Méthodologie

2.1 Extraction des points clés via MMPose

La première étape de notre travail consiste à numériser le mouvement humain au moyen de la librairie MMPose, sélectionnée pour sa robustesse et sa rapidité d'inférence. S'il existe des architectures plus puissantes et précises, telles que HRNet ou ViTPose, nous avons fait le choix de ne pas les mobiliser car notre environnement de travail repose exclusivement sur une exécution sur unité centrale (CPU), ce qui aurait rendu l'utilisation de modèles plus lourds techniquement difficile et excessivement lente pour nos analyses. Bien que notre inspiration initiale provienne du jeu Just Dance, nous avons choisi de ne pas utiliser directement les séquences originales du jeu car la présence de nombreux éléments visuels en arrière-plan risque de fausser la détection des points clés. Nous avons donc privilégié l'utilisation d'une vidéo brute où la danseuse évolue devant un fond neutre, ce qui nous permet d'exploiter de manière optimale l'outil MMPoseInferencer avec un modèle pré-entraîné sur le dataset COCO capable de localiser 17 articulations standards sur le corps humain. Pour chaque image de la séquence, le modèle détecte la présence humaine et génère en sortie une matrice de dimension (N,17,2), où N représente le nombre total d'images tandis que les autres dimensions correspondent aux articulations et à leurs coordonnées cartésiennes respectives (x,y).

Ce choix technique s'avère particulièrement pertinent pour l'analyse de la Danse de l'Égyptien puisque MMPose offre une précision élevée sur les membres supérieurs, notamment les épaules, les coudes et les poignets, qui constituent les éléments centraux de cette chorégraphie géométrique.

2.2 Normalisation et invariance d'échelle

L'un des défis majeurs de l'analyse vidéo réside dans la variabilité de la distance entre le sujet et la caméra, ce qui nécessite la mise en place d'un protocole de normalisation systématique pour permettre une comparaison rigoureuse entre la vidéo originale et nos propres enregistrements. Pour ce faire, nous procédons à un changement de repère en déplaçant l'origine du système de coordonnées au milieu de la ligne reliant les deux épaules, afin que la position des membres soit mesurée relativement au buste et non au cadre fixe de l'image. Nous complétons cette étape par une mise à l'échelle anatomique où chaque coordonnée est divisée par la largeur des épaules du sujet, ce qui permet d'éliminer les biais liés à la taille des individus ou à leur proximité avec l'objectif. Cette méthode garantit que nos analyses portent exclusivement sur la gestuelle et non sur les conditions de prise de vue, assurant ainsi une base de comparaison stable et objective pour l'étude des mouvements chorégraphiques.

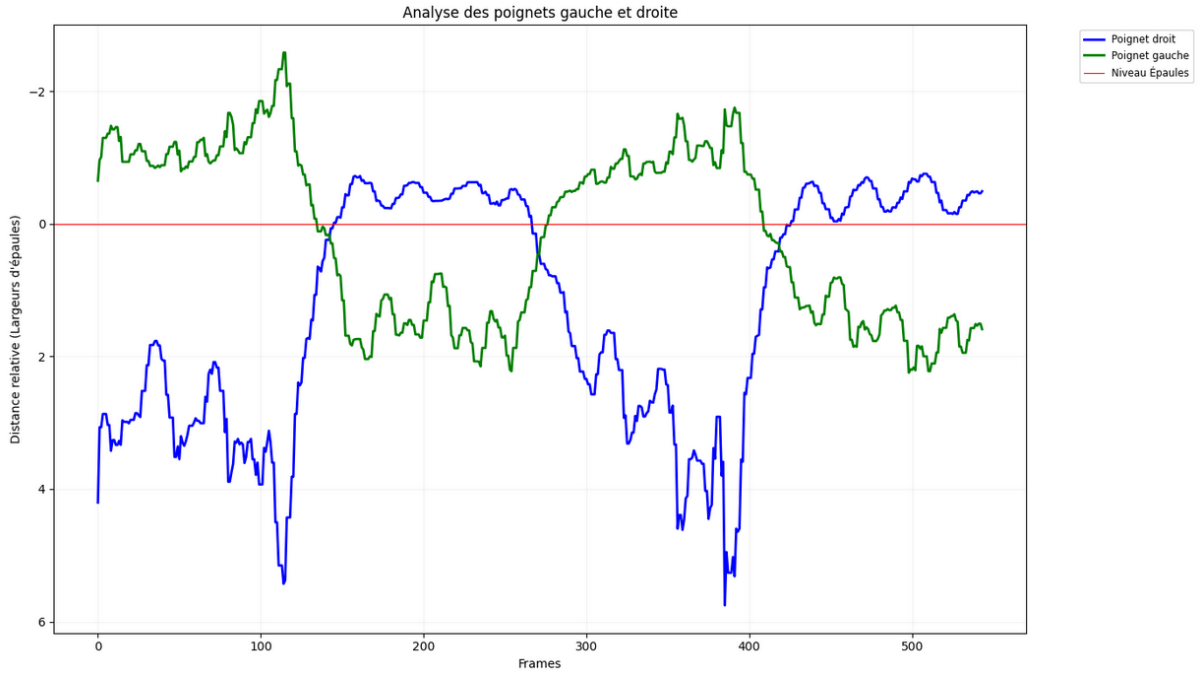

FIGURE 1 – Visualisation d'une pose générée : Le modèle respecte bien la géométrie angulaire typique de la danse.

2.3 Préparation des données pour l'apprentissage

Pour nos modèles de prédiction, nous structurons les données sous forme de fenêtres glissantes (sliding windows). Nous avons opté pour une fenêtre d'observation de 20 images pour prédire l'état de la 21ème, permettant ainsi au modèle de capturer la dynamique temporelle et le rythme de la danse.

3 Résultats et visualisations

3.1 Analyse Comparative : Référence vs Groupe

L'analyse de notre vidéo de groupe, impliquant quatre participants, nécessite une étape de traitement supplémentaire. Nous avons implémenté un algorithme de tri basé sur la position horizontale (coordonnée X) afin de suivre chaque individu de manière distincte tout au long de la séquence et d'éviter les inversions d'identités lors de l'extraction des trajectoires.

En superposant les courbes des poignets (gauche et droit) de la vidéo de référence avec celles de nos quatre danseurs, nous pouvons quantifier la fidélité de notre reproduction. Cette comparaison repose sur le calcul de l'erreur quadratique moyenne (MSE) entre les signaux normalisés. Nous observons ainsi les décalages de phase (problèmes de rythme) ou d'amplitude (problèmes d'extension des membres) par rapport au modèle original.

3.2 Modélisation et Prédiction par Réseaux Récurrents (LSTM)

Pour la génération autonome de mouvements, nous avons conçu un réseau de type LSTM (Long Short-Term Memory). Ce choix est motivé par la capacité de ces réseaux à mémoriser des dépendances temporelles à long terme, ce qui est indispensable pour respecter la structure cyclique d'une chorégraphie. Notre modèle se compose de deux couches cachées de 512 neurones avec un mécanisme de dropout pour prévenir le sur-apprentissage.

Le modèle a été entraîné sur 600 époques afin de minimiser l'écart entre les poses prédites et les poses réelles. Lors de la phase de génération, nous avons introduit un coefficient d'amplification α=1.5. Ce paramètre permet de "booster" les mouvements prédits, évitant ainsi que l'IA ne converge vers une pose statique moyenne et lui conférant une gestuelle plus dynamique et expressive.

La qualité de nos prédictions est évaluée par le score R^2 , qui mesure la corrélation entre les trajectoires générées par l'IA et les trajectoires réelles. Un score élevé atteste que le modèle a correctement assimilé les contraintes géométriques et temporelles de la danse de l'égyptien.

FIGURE 2 – Visualisation d'une pose générée : Le modèle respecte bien la géométrie angulaire typique de la danse.

4 Discussion et pistes d'amélioration

Malgré la pertinence des résultats obtenus, nous avons identifié une limite majeure résidant dans la « dérive anatomique » au cours de la génération prolongée. Sans contraintes physiques explicites, le réseau de neurones peut induire des déformations irréalistes des segments corporels, comme l'allongement anormal des membres. Pour pallier ce problème, une perspective d'amélioration consiste à explorer l'apprentissage par renforcement via l'algorithme PPO (Proximal Policy Optimization). Cette approche permettrait d'entraîner un agent virtuel à maintenir la posture égyptienne en recevant des récompenses basées sur la fidélité aux poses de référence tout en respectant des contraintes biomécaniques rigides. En conclusion, ce projet valide une chaîne complète de traitement du mouvement et démontre que la normalisation anatomique est une étape déterminante pour l'analyse objective de performances artistiques par l'intelligence artificielle.