Résumé en Français

Le mouvement humain a fasciné scientifiques, philosophes et artistes depuis l'Antiquité. L'avènement de la photographie et du cinéma, permettant de capturer les postures avec précision, a ouvert la voie à l'étude scientifique moderne du mouvement. Les systèmes d'intelligence artificielle (IA) peuvent désormais identifier la position des parties du corps humain à partir d'images, créant des représentations simplifiées en forme de "bonhomme allumette" correspondant à des points clés anatomiques, permettant une analyse automatique. Mais un seul point de vue ne fournit que des informations limitées : il est impossible de déterminer à quelle distance se trouvent les objets, et la vue peut être partiellement bloquée. En utilisant plusieurs vues de la même scène sous différents angles, il est possible d'obtenir des informations en 3D, de la même façon que nos deux yeux travaillent ensemble pour nous donner la perception de la profondeur. Cette thèse propose des méthodes pour concevoir des systèmes pratiques et robustes capables de percevoir et de comprendre les activités humaines et les scènes en 3D en utilisant plusieurs caméras.

Pour que plusieurs caméras puissent fonctionner ensemble, leurs positions et orientations respectives doivent être connues exactement. Cela peut se faire en *calibrant* les caméras, un processus qui requiert un équipement spécialisé, comme des motifs en damier, déplacés à travers la scène selon des procédures spécifiques, ce qui nécessite souvent un opérateur expert pour obtenir une bonne précision. Cela peut être long et coûteux, rendant impraticables les applications nécessitant une installation rapide. Dans cette thèse, nous présentons une nouvelle méthode de calibrage qui utilise les personnes se déplaçant dans la scène comme motifs de calibrage, permettant une installation plus rapide sans équipement: il suffit que quelqu'un se promène dans la scène pendant un moment.

Dans des conditions de laboratoire contrôlées, des caméras parfaitement calibrées permettraient une reconstruction parfaite des poses humaines en 3D. Mais les environnements réels présentent des défis importants. Les meubles, d'autres personnes, ou même l'environnement lui-même peuvent occulter des parties d'une personne, les conditions d'éclairage peuvent varier, et les estimations de posture initiales peuvent être bruitées ou inexactes. Simplement combiner ces observations imparfaites aboutit souvent à des postures 3D déformées ou absurdes. Pour reconstruire des poses humaines 3D précises même dans des conditions difficiles, nous avons entraîné un réseau de neurones à prêter attention aux relations géométriques entre les vues et à l'évolution des poses 2D au fil du temps. Il apprend à recréer des postures et mouvements humains 3D cohérents et précis même lorsque certaines parties du corps ne sont visibles par

aucune des caméras, comblant les lacunes avec des prédictions réalistes basées sur la façon dont les gens se déplacent naturellement.

Une fois que nous avons des séquences de mouvement 3D précises, nous pouvons enseigner aux ordinateurs à reconnaître différentes activités, comme "marcher" ou "faire des pompes". Entraîner des ordinateurs pour la reconnaissance d'activités nécessite habituellement de grandes quantités de données annotées, où des humains marquent et étiquettent manuellement chaque activité dans des milliers de séquences de mouvement. C'est intensif en main-d'œuvre et très coûteux. De plus, reconnaître une nouvelle activité nécessite de collecter et d'étiqueter de nouvelles données. D'autre part, simplement collecter des données de mouvement non étiquetées est facile et peu coûteux. Nous avons développé une méthode d'apprentissage auto-supervisé qui permet aux ordinateurs d'apprendre des représentations significatives à partir de grandes quantités de données de mouvement non étiquetées. Par exemple, elle permet d'apprendre qu'une action reste fondamentalement la même alors qu'elle est vue sous différents angles, ou lorsqu'elle est exécutée à différentes vitesses. Après cette phase d'apprentissage initial, appelée pré-entraînement, le système est très adaptable: il peut être entraîné à reconnaître de nouvelles activités en utilisant seulement quelques exemples étiquetés, réduisant considérablement le temps et le coût d'entraînement.

Comprendre le comportement humain nécessite plus que de simplement suivre le mouvement du corps. Nous interagissons constamment avec notre environnement, et nos mouvements prennent un sens spécifique selon le contexte: vous tenir debout devant un réfrigérateur suggère que vous allez l'ouvrir; vous diriger vers une chaise indique que vous êtes sur le point de vous asseoir. Sans voir les objets environnants, ces mouvements pourraient signifier presque n'importe quoi. Nous avons développé une méthode pour localiser des objets en 3D en utilisant seulement quelques points de vue ainsi que des descriptions en langage courant, comme "un bol bleu" ou "une chaise de bureau". Cela permet au système de trouver des objets qu'il n'a jamais été entraîné à reconnaître, fournissant un contexte environnemental riche pour l'activité humaine.

Ensemble, ces contributions aident à faire passer les systèmes de perception 3D multi-caméras du laboratoire au monde réel, les rendant plus accessibles, efficaces et adaptables à diverses applications telles que l'analyse sportive, la surveillance de sécurité, le divertissement, et au-delà.