OpenPose : Estimation de Pose 2D Multi-Personnes en Temps Réel utilisant les Champs d'Affinités par Partie

Papier original par Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, et Yaser Sheikh
Résumé en français par Cogoluègnes Charles et Le Berre Samuel

Objectifs

Le papier OpenPose a des objectifs multiples. Avant de les présenter, nous allons, dans un premier temps, parler des défis à relever. Par la suite nous évoquerons les approches existantes et enfin nous préciserons ces objectifs.

Figure 1 - Estimation de pose multi-personnes

Défis

Le système doit permettre de gérer des images avec un nombre de personnes non définies apparaissant à des positions et des tailles différentes. Il devient alors difficile d'associer les membres des personnes à cause des interactions, de l'environnement spatial et de l'occlusion. La complexité de calcul ne doit pas augmenter avec le nombre de personnes.

Approches existantes

La première approche s'intitule « Top Down ». L'idée de base est de faire une détection (ou estimation) pour chaque personne dans l'image. Cependant, le temps d'exécution est proportionnel au nombre de personnes. De plus, si la détection d'une personne échoue, (une erreur de proximité par exemple) alors on ne peut pas récupérer de données sur l'individu.

La seconde approche est intitulée "Bottom Up". Cette dernière évite les erreurs de détection de base et n'associe pas le temps d'exécution au nombre de personnes. Mais elle n'utilise pas le contexte global des autres parties du corps. Cela implique donc un calcul long à cause de l'inférence global sur la fin de ce dernier.

Buts

Pour résumer, le premier objectif est la détection d'un certain nombre de personnes dans une image. Le second est la prise en compte de n'importe quel type d'interaction entre ces personnes. Le dernier est d'avoir un temps de calcul qui n'est pas dépendant du nombre de personnes dans l'image.

Méthodes

Le but est d'effectuer une estimation de pose de multi-personnes avec une performance compétitive sur plusieurs benchmarks.

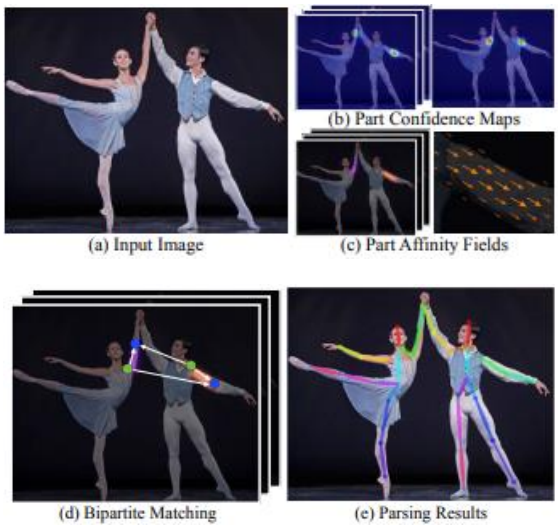

Figure 2 - Flux de tâches

Une méthode améliorée de "Bottom Up" (dite ascendante) a été retenue. Celle-ci utilise l'association des scores avec les "Champs d'Affinités par partie" (PAF), un jeu de vecteur de champs 2D qui représente la position et l'orientation des membres sur l'image.

Cette représentation ascendante (détection et association) montre un contexte global suffisant afin qu'une analyse profonde produise des résultats de haute qualité avec un coût de calcul moindre.

Architecture de réseau

Le réseau prédit de manière itérative les deux choses suivantes : les champs d'affinité d'association de membre à membre et les cartes de confiance de détection.

Cette prédiction itérative affine les résultats de la prédiction globale au cours de plusieurs étapes.

Alors que l'ancien modèle comportait des couches de convolution de taille 7x7, le nouveau lui comprend 3 noyaux consécutifs de taille 3x3. Ce changement réduit le nombre d'opérations par un facteur de ½. La sortie des 3 noyaux est concaténé en suivant l'approche de DenseNet. Le nombre de couches non linéaires est triplé et le réseau garde les fonctionnalités de bas et hauts niveaux.

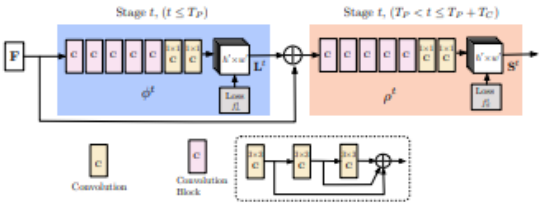

Figure 3 - Architecture à plusieurs niveaux CNN

Détection et association simultanées

L'image est analysée par un CNN qui génère une liste de traits qui est l'entrée de la première étape. Durant cette première étape le réseau produit un ensemble de champs d'affinité des membres.

À chaque étape suivante, les prédictions de l'étape précédente et les caractéristiques de l'image d'origine sont concaténées et utilisées afin de produire des prédictions affinées.

Après avoir bouclé sur le nombre total d'étapes PAF, ce processus est répété pour la détection des cartes de confiance, en commençant par la prédiction PAF la plus à jour.

Points importants

- PAF est crucial pour maximiser la précision, alors que les données du corps ne le sont pas. En gardant seulement les données des membres, la rapidité est augmentée de 200% et la précision de 7%.
- Un jeu de données de 15 000 pieds humains labellisés est présenté. L'utilisation de la reconnaissance des pieds avec le modèle n'entraîne pas un temps de calcul plus élevé, conserve la précision et améliore la détection du corps de manière générale.
- La méthode est généralisable (exemple d'utilisation avec des véhicules).

Figure 4 - Utilisation sur des véhicules

Résultats

Le premier système de reconnaissance multi-personnes en temps réel détectant le corps, les mains, les pieds et le visage.

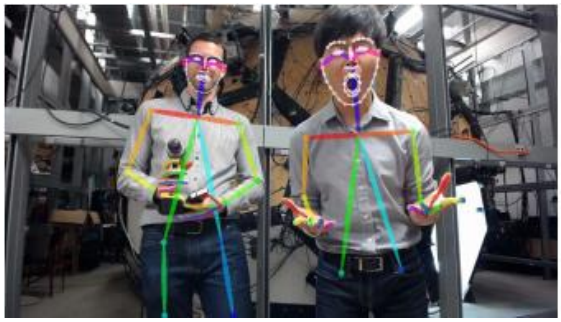

Figure 5 - Résultat des différents points de détection

Avantages

OpenPose peut fonctionner sur n'importe quel système d'exploitation et utiliser ou non l'accélération matérielle. Les entrées peuvent être une image, une vidéo, une webcam ou l'adresse IP d'une caméra. Il existe un très grand nombre de paramètres de haut niveau

(composante de détection, nombre de GPU, etc.) ce qui le rend très flexible et configurable. OpenPose a été entraîné sur les jeux de données COCO et MPII. La librairie comprend aussi un module de détection en 3D, possible grâce à une triangulation 3D avec un raffinement de Levenberg-Marquardt non linéaire sur les résultats de plusieurs caméras synchronisées. Le système est supérieur par rapport à d'autres méthodes récentes, telles que le Mask R-CNN ou AlphaPose (en prenant la ratio rapidité / précision).

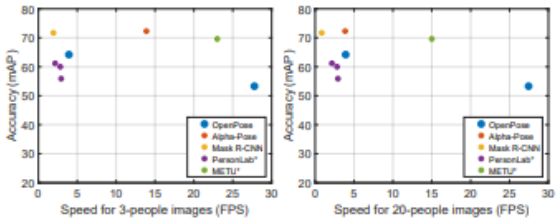

Figure 6 - Comparaison entre les différentes méthodes

OpenCV a intégré OpenPose en tant que module car il est très utilisé par les chercheurs et très efficace.

Limites

Avant OpenPose, les jeux de données ne décrivaient que les chevilles, le genou, la hanche, l'épaule, le coude, le poignet, le cou, le torse et le dessus de tête. Cependant, certains algorithmes ont besoin de données plus précises au niveau des pieds tels que le gros orteil ou le talon.

Pour répondre au problème, des données de pieds ont été extraites de COCO. Celles-ci ont pu être labellisées avec la plateforme Clickworker.

Au total 14 000 annotations en 3D ont été ajoutées avec environ 3 annotations par pied, pour créer un jeu de données de pieds.

Un modèle de détection de pieds aurait pu être utilisé dans OpenPose cependant cela aurait posé des problèmes comme ceux de l'approche "Top Down". C'est pour cela qu'il est intégré au détecteur de corps. De plus, la fusion des deux permet au détecteur de corps d'être plus précis, en particulier sur les jambes et chevilles.

On peut aussi observer des erreurs de détection, notamment lorsque le pied ou la jambe est occulté par le corps.

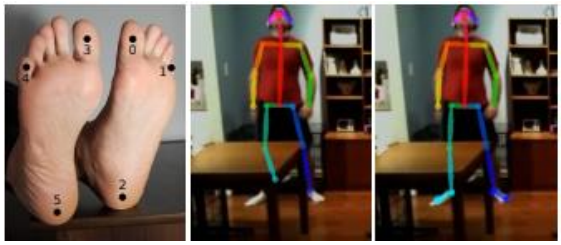

Figure 7 - Différence entre sans et avec les pieds

Conclusion

L'estimation de pose multi-personnes est un composant critique afin de permettre aux machines de mieux appréhender et comprendre le comportement humain avec leurs interactions. Dans ce papier, 4 buts ont été atteints :

- Avoir des résultats cohérents et de haute qualité.
- Pouvoir produire ces résultats dans un temps raisonnable.
- Que la complexité du système ne soit pas impactée (ou de très peu) par l'augmentation du nombre de personnes dans l'image.
- Rendre le code open source afin de laisser la possibilité à d'autre d'exploiter ce système pour proposer toujours plus d'applications innovantes.

Sources

[1] 2018_CVPR_Cao_OpenPose.pdf

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[3] https://www.learnopencv.com/multi-person-pose-estimation-in-opencv-using-openpose