Soutenance de Thèse

Institut de Recherche
en Informatique de Toulouse

Jeudi 19 Septembre 2019
10h00
UT3 Paul Sabatier, IRIT, Auditorium J. Herbrand

Huy-Hieu PHAM
Equipe MINDS, IRIT

Architectures d'apprentissage profond pour la reconnaissance d'actions humaines dans des séquences vidéo RGB-D monoculaires. Application à la surveillance dans les transports publics

Jury :
- M. Denis KOUAME – *Université Toulouse III - Paul Sabatier, directeur de thèse*
- M. Stéphane CANU – *INSA Rouen, rapporteur*
- M. Yassine RUICHEK – *Université de Technologie de Belfort-Montbéliard, examinateur*
- Mme Michèle GOUIFFÈS – *Université Paris Sud, rapporteur*
- M. Louahdi KHOUDOUR – *CEREMA, co-directeur de thèse*
- M. Frédéric LERASLE – *Université Toulouse III - Paul Sabatier, examinateur*
- M. Alain CROUZIL – *Université Toulouse III - Paul Sabatier, encadrant, invité*
- M. Jean-Marc DAVIAU – *Tisséo, invité*
- M. Sergio A VELASTIN – *Cortexica Vision Systems Ltd. Londres, invité*

Mots-clés : reconnaissance d'actions humaines, réseaux de neurones convolutifs, recherche d'architecture neuronale, squelettes, capteur de profondeur

Résumé : Cette thèse porte sur la reconnaissance d'actions humaines dans des séquences vidéo RGB-D monoculaires. La question principale est, à partir d'une vidéo ou d'une séquence d'images donnée, de savoir comment reconnaître des actions particulières qui se produisent. Cette tâche est importante et est un défi majeur à cause d'un certain nombre de verrous scientifiques induits par la variabilité des conditions d'acquisition, comme l'éclairage, la position, l'orientation et le champ de vue de la caméra, ainsi que par la variabilité de la réalisation des actions, notamment de leur vitesse d'exécution. Pour surmonter certaines de ces difficultés, dans un premier temps, nous examinons et évaluons les techniques les plus récentes pour la reconnaissance d'actions dans des vidéos. Nous proposons ensuite une nouvelle approche basée sur des réseaux de neurones profonds pour la reconnaissance d'actions humaines à partir de séquences de squelettes 3D. Deux questions clés ont été traitées. Tout d'abord, comment représenter la dynamique spatio-temporelle d'une séquence de positions du squelette pour exploiter efficacement la capacité d'apprentissage des représentations de haut niveau des réseaux de neurones convolutifs (CNNs ou ConvNets). Ensuite, comment concevoir une architecture de CNN capable d'apprendre des caractéristiques spatio-temporelles discriminantes à partir de la représentation proposée dans un objectif de classification. Pour cela, nous introduisons deux nouvelles représentations du mouvement 3D basées sur des squelettes, appelées SPMF (Skeleton Posture-Motion Feature) et Enhanced-SPMF, qui encodent les postures et les mouvements humains extraits des séquences de squelettes sous la forme d'images couleur RGB. Pour les tâches d'apprentissage et de classification, nous proposons différentes architectures de CNNs, qui sont basées sur les modèles

05 61 55 65 10
info@irit.fr
www.irit.fr

Institut de Recherche
en Informatique de Toulouse

Residual Network (ResNet), Inception-ResNet-v2, Densely Connected Convolutional Network (DenseNet) et Efficient Neural Architecture Search (ENAS), pour extraire des caractéristiques robustes de la représentation sous forme d'image que nous proposons et pour les classer. Les résultats expérimentaux sur des bases de données publiques (MSR Action3D, Kinect Activity Recognition Dataset, SBU Kinect Interaction, et NTU-RGB+D) montrent que notre approche surpasse les méthodes de l'état de l'art, tout en nécessitant de faibles temps d'exécution pour l'apprentissage et pour l'inférence. Nous proposons également une nouvelle technique pour l'estimation de postures humaines à partir d'une vidéo RGB. Pour cela, le modèle d'apprentissage profond appelé OpenPose est utilisé pour détecter les personnes et extraire leur posture en 2D. Un réseau de neurones est ensuite proposé pour apprendre la transformation permettant de reconstruire ces postures en trois dimensions. Les résultats expérimentaux sur la base de données Human3.6 montrent l'efficacité de la méthode proposée. Ces résultats ouvrent des perspectives pour une approche de la reconnaissance d'actions humaines à partir des séquences de squelettes 3D sans utiliser des capteurs de profondeur comme la Kinect. Nous avons également constitué CEMEST, un nouveau jeu de données RGB-D illustrant des comportements de passagers dans les transports publics. Il contient 203 vidéos de surveillance collectées dans une station du métro incluant des événements « normaux » et « anormaux ». Nous avons obtenu des résultats prometteurs dans des conditions réelles en utilisant des techniques d'augmentation de données et de transfert d'apprentissage. Notre approche permet de concevoir des applications réelles basées sur des techniques de l'apprentissage profond pour renforcer la qualité des services de transport en commun.

05 61 55 65 10
info@irit.fr
www.irit.fr