Soutenance de thèse : Thi Hai Hong Phan

10 September 2019, 13:30

Titre de la thèse

Reconnaissance d'actions humaines dans des vidéos avec l'apprentissage automatique.

Date et lieu de soutenance

Mardi 10 septembre 2019, 13h30.

ENSEA Cergy, salle du conseil.

Résumé

Ces dernières années, la reconnaissance d'action humaine (HAR) a attiré l'attention de la recherche grâce à ses diverses applications telles que les systèmes de surveillance intelligents, l'indexation vidéo, l'analyse des activités humaines, les interactions homme-machine, et ainsi de suite. Les problèmes typiques que les chercheurs envisagent sont la complexité des mouvements humains, les variations spatio-temporelles, l'encombrement, l'occlusion et le changement des conditions d'éclairage. Cette thèse porte sur la reconnaissance automatique des actions humaines en cours dans une vidéo. Nous abordons ce problème de recherche en utilisant à la fois des approches d'apprentissage traditionnel peu profond et d'apprentissage profond.

Premièrement, nous avons commencé les travaux de recherche avec des méthodes d'apprentissage traditionnelles peu profondes, fondées sur des caractéristiques créées manuellement, en introduisant un nouveau fonctionnalité appelée descripteur MOMP (Motion of Oriented Magnitudes Patterns). Nous avons ensuite intégré ce descripteur discriminant aux techniques de représentation simples mais puissantes telles que le sac de mots visuels, le vecteur de descripteurs agrégés localement (VLAD) et le vecteur de Fisher pour mieux représenter les actions. En suite l'PCA (Principal Component Analysis) et la sélection des caractéristiques (la dépendance statistique, l'information mutuelle) sont appliquées pour rechercher le meilleur sous-ensemble des caractéristiques afin d'améliorer les performances et de réduire les coûts de calcul. La méthode proposée a permis d'obtenir les résultats d'état de l'art sur plusieurs bases de données communes.

Les approches d'apprentissage profond récentes nécessitent des calculs intensifs et une utilisation importante de la mémoire. Ils sont donc difficiles à utiliser et à déployer sur des systèmes aux ressources limitées. Dans la deuxième partie de cette thèse, nous présentons un nouvel algorithme efficace pour compresser les modèles de réseau de neurones convolutionnels afin de réduire à la fois le coût de calcul et l'empreinte mémoire au moment de l'exécution. Nous mesurons la redondance des paramètres en fonction de leurs relations à l'aide des critères basés sur la théorie de l'information, puis nous éliminons les moins importants. La méthode proposée réduit considérablement la taille des modèles de différents réseaux tels

qu'AlexNet, ResNet jusqu'à 70% sans perte de performance pour la tâche de classification des images à grande échelle.

L'approche traditionnelle avec le descripteur proposé a permis d'obtenir d'excellentes performances pour la reconnaissance de l'action humaine mais seulement sur de petits bases de données. Afin d'améliorer les performances de la reconnaissance sur les bases de données de grande échelle, dans la dernière partie de cette thèse, nous exploitons des techniques d'apprentissage profond pour classifier les actions. Nous introduisons les concepts de l'image MOMP en tant que couche d'entrée de CNN et incorporons l'image MOMP dans des réseaux de neurones profonds. Nous appliquons ensuite notre algorithme de compression réseau pour accélérer et améliorer les performances du système. La méthode proposée réduit la taille du modèle, diminue le sur-apprentissage et augmente ainsi la performance globale de CNN sur les bases de données d'action à grande échelle.

Tout au long de la thèse, nous avons montré que nos algorithmes obtenaient de bonnes performances sur bases de données d'action complexes (Weizmann, KTH, UCF Sports, UCF-101 et HMDB51) avec des ressources limitées.

Abstract

In recent years, human action recognition (HAR) has attracted the research attention thanks to its various applications such as intelligent surveillance systems, video indexing, human activities analysis, human-computer interactions and so on. The typical issues that the researchers are envisaging can be listed as the complexity of human motions, the spatial and temporal variations, cluttering, occlusion and change of lighting condition. This thesis focuses on automatic recognizing of the ongoing human actions in a given video. We address this research problem by using both shallow learning and deep learning approaches.

First, we began the research work with traditional shallow learning approaches based on hand-scrafted features by introducing a novel feature named Motion of Oriented Magnitudes Patterns (MOMP) descriptor. We then incorporated this discriminative descriptor into simple yet powerful representation techniques such as Bag of Visual Words, Vector of locally aggregated descriptors (VLAD) and Fisher Vector to better represent actions. Also, PCA (Principal Component Analysis) and feature selection (statistical dependency, mutual information) are applied to find out the best subset of features in order to improve the performance and decrease the computational expense. The proposed method obtained the state-of-the-art results on several common benchmarks.

Recent deep learning approaches require an intensive computations and large memory usage. They are therefore difficult to be used and deployed on the systems with limited resources. In the second part of this thesis, we present a novel efficient algorithm to compress Convolutional Neural Network models in order to decrease both the computational cost and the run-time memory footprint. We measure the redundancy of parameters based on their relationship using the information theory based criteria, and we then prune the less important ones. The proposed method significantly reduces the model sizes of different networks such as AlexNet, ResNet up to 70% without performance loss on the large-scale image classification task.

Traditional approach with the proposed descriptor achieved the great performance for human action recognition but only on small datasets. In order to improve the performance on the large-scale datasets, in the last part of this thesis, we therefore exploit deep learning techniques to classify actions. We introduce the concepts of MOMP Image as an input layer of CNNs as well as incorporate MOMP image into deep neural networks. We then apply our network compression algorithm to accelerate and improve the performance of system. The proposed method reduces the model size, decreases the over-fitting, and thus increases the overall performance of CNN on the large-scale action datasets.

Throughout the thesis, we have showed that our algorithms obtain good performance in comparison to the

state-of-the-art on challenging action datasets (Weizmann, KTH, UCF Sports, UCF-101 and HMDB51) with low resource required.

Mots-clefs

reconnaissance d'actions, Classification, apprentissage automatique, compression profond

Keywords

Action Recognition, classification, shallow learning, deep learning, deep compression

Composition du jury

- Mathias QUOY, Professeur des Universités, Université de Cergy-Pontoise, Directeur de thèse
- Catherine ACHARD, Maîtres de conférences, HDR, Université Pierre et Marie Curie, Rapporteur
- Alice CAPLIER, Professeur des Universités, Grenoble-INP Institut d'ingénierie Univ. Grenoble Alpes, Rapporteur
- Antoine MANZANERA, Professeur, ENSTA ParisTech - École Nationale Supérieure de Techniques Avancées, Examinateur
- Thanh Phuong NGUYEN, Maître de Conférences, Université de Toulon, Examinateur
- Ngoc Son VU, Maître de Conférences, ENSEA-École Nationale Supérieure de l'Electronique et de ses Applications, Co-encadrant de thèse