Sujet de thèse - campagne 2017

Laboratoire : Institut des Systèmes Intelligents et de Robotique (ISIR)

Etablissement de rattachement : UPMC

Titre de la thèse : Apprentissage automatique pour l'analyse des expressions faciales en environnement non contraint

Directeur de thèse : Mohamed CHETOUANI (PU, CNU61)

Codirection : Kévin BAILLY (MCF, CNU61)

Mail de contact : mohamed.chetouni@upmc.fr, kevin.bailly@upmc.fr

Collaborations dans le cadre de la thèse :

Rattachement à un programme :

Cotutelle envisagée :

Si oui avec quelle université & quel laboratoire :

Le sujet peut-il être publié sur le site web de l'ED SMAER : oui

Résumé du sujet :

Le visage d'un être humain est constitué d'une quarantaine de muscles qui permettent de produire quelques milliers d'expressions faciales. Ces expressions sont porteuses d'informations relatives à l'état cognitif et aux intentions sociales d'une personne. L'analyse des expressions faciales vise à extraire de manière automatique ces indices dans une image ou un flux vidéo. Il s'agit d'un domaine de recherche très actif à l'interface de la vision par ordinateur et de l'apprentissage artificiel.

L'objectif scientifique de cette thèse est de **concevoir des méthodes d'apprentissage à l'interfaces des réseaux de neurones profonds et des forêts aléatoire** (également appelé réseaux conditionnels) afin de **concevoir des systèmes d'analyse d'expressions faciales en conditions réelles**. Les méthodes proposées devront donc être robustes aux différentes variations de l'environnement (illumination, angle de vue, identité de la personne) et temps réel (contexte interactif)

Les modèles d'analyse d'expressions faciales développés durant cette thèse seront diffusés sous licence Open Source afin d'être intégrés dans de nouvelles interfaces utilisateur naturelles ou des robots sociaux.

Sujet développé

Contexte

Les expressions faciales sont porteuses d'informations relatives à l'état cognitif, émotionnel et aux intentions sociales d'une personne. La reconnaissance des expressions faciales vise à extraire de manière automatique ces indices dans une image ou un flux vidéo. Il s'agit d'un domaine de recherche très actif à l'interface de la vision par ordinateur et de l'apprentissage statistique. Depuis une quinzaine d'années, les recherches sur l'analyse automatique des visages ont connu des avancées formidables et les technologies actuelles sont suffisamment matures pour être exploitées dans des applications et des services commerciaux tels que la capture de mouvements, le marketing et les études médicales et comportementales (objectivation de troubles cognitifs).

Toutefois, pour être vraiment effectifs et déployés à plus large échelle, les systèmes d'analyse faciale de nouvelle génération devront répondre à plusieurs défis d'envergure :

(a) **La grande variabilité des données en conditions réelles (*in the wild*)** : les expressions faciales sont très variables dans leur dynamique et leur intensité, et l'apparence d'un visage est influencée par de nombreux facteurs indépendants de l'expression tels que la pose, l'identité du sujet, son âge, son ethnie ou encore les conditions d'illumination, les occultations et le type de camera utilisé.

(b) **Le faible nombre de données annotées** : Dans le domaine de l'analyse des expressions faciales, les risques de surapprentissage sont particulièrement importants car le nombre de données annotées disponibles pour entrainer les modèles prédictifs est limité. L'annotation manuelle d'images de visage est fastidieuse et certaines tâches telles que la détection des activations musculaires (*Action Units*) requièrent une expertise spécifique. En effet il s'agit de mouvements faciaux souvent subtils qui ne peuvent être annotés que par des personnes formées et certifiées.

(c) **La complexité mémoire des modèles** : ces modèles devront être compacts afin d'être embarqués sur un robot mobile ou sur un téléphone portable.

(d) **La complexité en temps des modèles** : dans un contexte d'interaction homme machine il est essentiel que l'analyse s'effectuent en temps réel.

(e) **L'adaptabilité des modèles à de nouveaux domaines et à de nouvelles tâches** : les méthodes actuelles ont une capacité de généralisation limitée lorsque les données de test diffèrent sensiblement de celles d'apprentissage. De même, certaines tâches à réaliser peuvent être différentes mais fortement corrélées (la reconnaissance des émotions et la détection des activations musculaires du visage par exemple). Adapter un modèle consiste alors à exploiter les connaissances d'un domaine et d'une tâche source afin de traiter un nouveau domaine et une nouvelle tâche cible.

Objectifs

Pour relever ces défis, le **choix du modèle prédictif, de son architecture, et de l'algorithme d'apprentissage est donc central**. Depuis quelques années, **les approches par réseaux de neurones profonds ont connu un immense succès dans de nombreuses tâches d'apprentissage**, et en particulier en analyse d'image. Ils sont capables d'intégrer l'information issue de très grandes bases de données (plusieurs millions d'images) et d'apprendre conjointement un espace de représentation des données et des fonctions de prédictions complexes. De plus, la nature différentiable de ces modèles permet d'ajuster facilement les paramètres d'un réseau lorsque l'on dispose de nouvelles données d'apprentissage. Ils sont donc bien adaptés pour répondre aux défis (a) et (e) mais présentent deux inconvénients majeurs : les temps de prédictions peuvent être longs, en particulier lors de l'évaluation des couches complètement connectées et les réseaux sont très sensibles au surapprentissage lorsque le nombres de données est faible au regard de la tâche à prédire.

Parallèlement à ces modèles, les méthodes d'ensemble telles que les forêts aléatoires constituent une alternative intéressante car les temps d'exécution sont faibles (pour chaque arbre seul un chemin est évalué) et la combinaison des prédictions des modèles appris sur des échantillons de données différents limite les risques de surapprentissage. Ils répondent donc bien aux défis (b) et (d). Par ailleurs, l'équipe a déjà proposé des méthodes d'analyse d'expressions

faciales qui s'appuient des forêts aléatoires pour gérer les problèmes de robustesse aux occultations [1] et aux variations de poses [2] et de dynamique [3] (défi (a)). Cependant, les modèles sont souvent volumineux (plusieurs centaines d'arbres à stocker en mémoire) et non différentiables, ce qui ne permet pas d'adapter les modèles appris à de nouveaux contextes.

Très récemment, **des modèles hybrides également appelés modèles conditionnel [4] ont cherché à combiner les avantages des réseaux de neurones profonds et des forêts aléatoires** et offrent des perspectives de recherche très intéressantes [5], [6], [7].

L'objectif scientifique de cette thèse est donc de **proposer de nouvelles architectures de réseaux hybrides permettant de repousser les limites actuelles des réseaux conditionnels afin d'élaborer de nouveaux systèmes d'analyse faciale robustes (défis a et b), légers (défi c), temps réels (défi d), et adaptables (défi e).**

Résultats attendus

La première partie de la thèse consistera à **concevoir une méthode de localisation des points caractéristiques robuste à la pose et aux occultations** qui constituent les deux sources de variations les plus difficiles à gérer par les systèmes actuels. La nature différentiable des modèles permettra également de tester les capacités d'adaptation des modèles sur un visage en particulier (via une étape de calibrage par exemple)

La seconde partie de la thèse sera consacrée à la **réalisation d'un système robuste et temps réel d'estimation de l'intensité des activations musculaires (Action Units)** qui s'appuiera sur la méthode de localisation des points caractéristiques développée précédemment.

Les modèles d'analyse d'expressions faciales développés durant cette thèse seront évaluées dans des **cas d'utilisation en environnement non contraint** et ont vocation à être diffusés sous **licence Open Source** afin d'être intégrés dans de nouvelles interfaces utilisateur naturelles ou des robots sociaux.

Références bibliographiques

[1] A. Dapogny, K. Bailly, and S. Dubuisson, "**Confidence-Weighted Local Expression Predictions for Occlusion Handling in Expression Recognition and Action Unit detection**," International Journal of Computer Vision (IJCV), 2017

[2] A. Dapogny, K. Bailly, and S. Dubuisson, "**Dynamic Pose-Robust Facial Expression Recognition by Multi-View Pairwise Conditional Random Forests**," in arXiv:1607.06250, 2016

[3] A. Dapogny, K. Bailly, and S. Dubuisson, "**Pairwise Conditional Random Forests for facial expression recognition**," in International Conference on Computer Vision (ICCV), 2015

[4] Y. Ioannou, D. Robertson, D. Zikic, P. Kontschieder, J. Shotton, M. Brown, A. Criminisi, "**Decision forests, convolutional networks and the models in-between**," in arXiv:1603.01250

[5] P. Kontschieder, M. Fiterau, A. Criminisi, S. Bulo, "**Deep Neural Decision Forests**," in IEEE International Conference on Computer Vision (ICCV), 2015

[6] S. Rota Bulo et P. Kontschieder, « **Neural Decision Forests for Semantic Image Labelling** », in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, p. 81□88.

[7] A. Roy et S. Todorovic, « **Monocular Depth Estimation Using Neural Regression Forest** », in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, p. 5506□551