AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

M^me^ : FILALI Hajar

***Soutiendra** : le 17/12/2022 à 10H*
***Lieu** : Nouveau Centre Polyvalent des Etudes Doctorales de l'Université Sidi Mohamed Ben Abdellah - Amphi 1.*

***Une thèse intitulée** :*
Deep Meaningful Learning Based Hybrid Approaches for Human Emotion Recognition.

En vue d'obtenir le Doctorat
FD** : Sciences et Technologies de l'Information et de Communication **STIC
***Spécialité** : Informatique*

Devant le jury composé comme suit :

Nom et prénom	Etablissement	Grade	Qualité
EL BEQQALI Omar	Faculté des Sciences Dhar El Mahraz - Fès	PES	Président
OUANAN Mohamed	Faculté des Sciences-Meknès	PES	Rapporteur
YAHYAOUY Ali	Faculté des Sciences Dhar El Mahraz– Fès	PH	Rapporteur
EN-NAHNAHI Noureddine	Faculté des Sciences Dhar El Mahraz - Fès	PH	Rapporteur
BELLACH Benaissa	Ecole Nationale des Sciences Appliquées Oujda	PES	Examinateur
MAHRAZ Mohamed Adnane	Faculté des Sciences Dhar El Mahraz – Fès	PH	Examinateur
EL FAZAZY Khalid	Faculté des Sciences Dhar El Mahraz – Fès	PH	Examinateur
RIFFI Jamal	Faculté des Sciences Dhar El Mahraz – Fès	PH	Co-Directeur de thèse
TAIRI Hamid	Faculté des Sciences Dhar El Mahraz - Fès	PES	Directeur de thèse

Résumé :

La reconnaissance des émotions est devenue l'un des sujets les plus traités par la communauté scientifique. Un large éventail de caractéristiques peut être utilisé, telles que l'expression faciale, le langage corporel, le son, etc. Plusieurs approches ont été développées dans le but de renforcer la machine et de la doter de la capacité pour déchiffrer et lire les émotions des personnes afin d'interagir plus intelligemment. Cependant, cette tâche reste un grand défi autour duquel se rassemblent différentes communautés (interaction homme-machine, traitement d'image, l'intelligence artificielle, robotique, ..., etc). Des décennies de recherche scientifique ont été menées sur l'analyse unimodale des émotions, typiquement l'expression faciale, puisqu'elle représente 55\% de la communication non verbale qui permet de comprendre l'émotion d'une personne. Malgré une certaine amélioration de la précision pour les systèmes unimodal , la plupart des nouvelles approches se sont basées sur des techniques de classification de caractéristiques multimodales. L'utilisation de techniques d'apprentissage profond, pour extraire automatiquement des caractéristiques efficaces à partir d'informations unimodales ou multimodales, ainsi que leur utilisation dans la fusion et les classifications, sont actuellement des nouvelles directions poursuivies activement par les chercheurs, mais plusieurs défis restent à relever pour réaliser un système d'apprentissage profond et robuste pour prédire les émotions humaines.

Dans ce contexte, cette thèse de doctorat aborde les questions ci-dessus pour résoudre le problème de la reconnaissance automatique du comportement émotionnel humain et améliorer la précision de la reconnaissance des émotions unimodales et multimodales. En proposant une nouvelle architecture de réseau de neurones appelée réseau de neurones significatifs, cette dernière permet de classer de manière significative différentes modalités. En effet, notre réseau apprend les composantes de chaque vecteur séparément et applique une concaténation jusqu'à la couche de fusion, contrairement aux autres architectures qui fusionnent les modalités sans prendre en compte les différentes composantes du vecteur résultant pour chacune d'entre elles. L'approche proposée s'est d'abord basée sur une architecture hybride d'apprentissage profond pour la reconnaissance des émotions faciales utilisant le réseau de neurones convolutif et l'autoencodeur empilé. L'idée principale de ce travail est de combiner les deux vecteurs caractéristiques générés par chacune de ces architectures et d'alimenter le vecteur résultant au réseau neuronal significatif. L'architecture de ce dernier conduit à apprendre chaque caractéristique en dédiant un ensemble de neurones pour chaque composante du vecteur avant de les combiner tous ensemble dans la dernière couche. Cette approche est testée sur quatre bases de données publiques et une étude comparative est établie pour garantir et justifier l'efficacité et la robustesse de cette méthode. Dans un deuxième temps, nous exploitons l'efficacité de réseau de neurone significatif pour permettre la prédiction multimodal des émotions pendant une conversation. En utilisant les modalités texte et audio, nous avons proposé des méthodes d'extraction de caractéristiques basées sur le l'apprentissage profond. Ensuite, nous utilisons la modalité bimodale qui est créée suite à la fusion des caractéristiques textuelles et audio. Les vecteurs de caractéristiques de ces trois modalités sont affectés à l'alimentation du réseau neuronal significatif afin d'apprendre séparément chaque caractéristique. Ce modèle a été évalué sur l'ensemble de données multimodales et multipartites pour la reconnaissance des émotions dans la conversation MELD. L'approche proposée a atteint une précision élevée qui surpasse de manière significative tous les systèmes multimodaux actuels.

Mots clés :

l'apprentissage profond, Reconnaissance des expressions faciales, Architecture hybride, Réseau neuronal convolutif, Stacked AutoEncoder. Réseau Neuronal Significatif, reconnaissance multimodale des émotions.

Deep Meaningful Learning Based Hybrid Approaches for Human Emotion Recognition

Abstract :
Emotion recognition has become one of the most researched subjects in the scientific community, a wide range of features can be used such as facial expression, body language, audio, etc. Several approaches have been developed aiming at strengthening the machine and endowing it, with the ability to decipher and read people's emotions to interact more intelligently. However, this task still a big challenge around which gather different communities (human-computer interaction, image processing, artificial intelligence, robotics, …, etc). Decades of scientific research have been conducted on unimodal emotion analysis, typically facial expression, since it represents 55\% of the nonverbal communication that allows to understand the emotion of a person. Despite some improvement of accuracy, most of the early approaches have relied on multimodal features classification techniques. The use of deep learning techniques to automatically extract effective features from unimodal or multimodal information as well as their use in fusion and classifications are new directions currently actively pursued by researchers, but several challenges remain to realize a deep and robust learning system to predict human emotions.
In this context, this PhD thesis address the above issues to tackle the problem of machine understanding concerning human affective behavior and improve the accuracy of both unimodal and multimodal emotion recognition. By proposing a new architecture of neural network called meaningful neural network, this last allows different modalities to be classified significantly. In fact, our network learns each vector's component separately and applies a concatenation until the fusion layer, not like the other architectures that merge the modalities without taking into consideration the different components of the resulting vector for each one. The proposed approach was based firstly on hybrid deep learning architecture for facial emotion recognition using Convolutional Neural Network and Stacked AutoEncoder. The main idea of this work is to combine the two feature vectors generated by each of these architectures and feed the resulting vector to the meaningful neural network. The architecture of the latter leads to learn each feature by dedicating a set of neurons for each component of the vector before combining them all together in the last layer. This approach is tested on four public databases and a comparative study is established to guarantee and justify the effectiveness and robustness of this method. Secondly, we exploit Meaningful Neural Network Effectiveness to enable emotion prediction during a conversation. Using the text and the audio modalities, we proposed feature extraction methods based on Deep Learning. Then, the bimodal modality that is created following the fusion of the text and audio features is used. The feature vectors from these three modalities are assigned to feed the meaningful neural network in order to separately learn each characteristic. This model was evaluated on a multimodal and multiparty dataset for emotion recognition in conversation MELD. The proposed approach reached a high accuracy which significantly outperforms all current multimodal systems.

Key Words :
Deep Learning, Facial expression recognition, Hybrid architecture, Convolutional neural network, Stacked AutoEncoder. Meaningful Neural Network, multimodal emotion recognition.