Nom et Prénom : KAS MOHAMED
Date de soutenance : le Mercredi 07 Juillet 2021 à 09h30 à la Faculté des Sciences Kénitra
Directeur de Thèse : ROCHDI MESSOUSSI

Sujet de Thèse :

Développement de méthodes à base de descripteurs locaux et apprentissage profond pour la reconnaissance du visage et des émotions faciales

Résumé :

Cette thèse porte sur le développement de nouveaux concepts de segmentation d'images et de classification de régions pour l'analyse d'images. Il s'agit de mettre en œuvre de nouveaux descripteurs, qu'ils soient de couleur, de texture ou de forme et proposer de nouvelles architectures d'apprentissage profond pour des applications liées à l'analyse faciale. Nous nous concentrons sur la reconnaissance faciale et la classification des expressions faciales. Notre thèse a débouché sur la proposition de nombreuses contributions liées à l'analyse faciale couvrant des architectures classiques et profondes. Nous avons contribué à la reconnaissance faciale tout d'abord par la proposition d'un descripteur local appelé Mixed Neighborhood Topology Cross Decoded Patterns. Notre descripteur de visage repose sur une nouvelle topologie de voisinage et une fonction de noyau avancée permettant en encodage efficace des caractéristiques liées à la personne. Nous avons évalué le système de reconnaissance faciale proposé à base du MNTCDP sur des bases de données connues de l'état de l'art, disposant de challenges, couvrant la diversité des individus, environnement non contrôlé, des conditions de fond et d'éclairage variables. Les résultats obtenus ont dépassé plusieurs résultats de l'état de l'art. Pour la deuxième contribution, nous avons relevé le défi de la reconnaissance faciale invariante aux poses (PIFR) en développant une méthode de génération d'images basée sur le Generative Adversarial Network, afin de générer une image frontale correspondant à une image en profil. Cette transformation rend la reconnaissance beaucoup plus facile puisque la plupart des bases de données de référence n'incluent que des échantillons de face frontale. Nous avons créé une architecture profonde de bout en bout, composé du GAN pour la génération des échantillons de profil et un classificateur basé sur ResNet pour l'identification de la personne à partir de son image frontale synthétisée. Les expériences, que nous avons menées sur une base de données adéquate, notamment en termes de chevauchement des individus entre la base de l'apprentissage et celle de l'évaluation, mettent en évidence une grande amélioration des performances du PIFR grâce à la génération d'images frontales du GAN. Nos contributions à la tâche de reconnaissance des expressions faciales (FER) couvrent à la fois des scénarios statiques (une image) et dynamiques (plusieurs images). Notre architecture pour FER avec le mode statique repose sur l'extraction de caractéristiques de texture et de forme à partir de points de repère du visage spécifiques qui présentent suffisamment d'informations pour détecter l'émotion dominante. Nous avons proposé un descripteur appelé Orthogonal and Parallel-based Directions Generic Query Map Binary Patterns pour extraire efficacement les caractéristiques texturales liées aux émotions à partir de 49 points de repère. Ces caractéristiques sont combinées avec celles à base de forme calculées à l'aide du descripteur HOG sur un masque binaire

représentant l'interpolation des 49 points. La classification est réalisée via le SVM. Less performances obtenues sur cinq bases de données avec le protocole Leave One Subject Out ont démontré l'efficacité de l'architecture proposée par rapport à l'état de l'art. D'autre part, notre contribution relative à la FER avec le mode dynamique intègre un réseau LSTM pour encoder avec précision les informations temporelles avec un masque d'attention permettant de se concentrer sur les repères liés aux émotions et garantir la robustesse de la reconnaissance. Nous avons considéré quatre échantillons comme entrées représentant l'évolution de l'émotion jusqu'à son pic. Chaque échantillon est codé via une branche CNN et les quatre branches sont jointes par un bloc LSTM qui prédit l'émotion dominante. Les expériences menées sur trois bases de données pour FER dynamique ont montré que l'architecture CNN-LSTM profonde proposée dépasse l'état de l'art.

Abstract :

The research objectives of this thesis concern the development of new concepts for image segmentation and region classification for image analysis. This involves implementing new descriptors, whether color, texture, or shape, to characterize regions and propose new deep learning architectures for the various applications linked to facial analysis. We restrict our focus on face recognition and person-independent facial expressions classification tasks, which are more challenging, especially in unconstrained environments. Our thesis lead to the proposal of many contributions related to facial analysis based on handcrafted and deep architecture. We contributed to face recognition by an effective local features descriptor referred to as Mixed Neighborhood Topology Cross Decoded Patterns (MNTCDP). Our face descriptor relies on a new neighborhood topology and a sophisticated kernel function that help to effectively encode the person-related features. We evaluated the proposed MNTCDP-based face recognition system according to well-known and challenging benchmarks of the state-of-the-art, covering individuals' diversity, uncontrolled environment, variable background and lighting conditions. The achieved results outperformed several state-of-the-art ones. As a second contribution, we handled the challenge of pose-invariant face recognition (PIFR) by developing a Generative Adversarial Network (GAN) based image translation to generate a frontal image corresponding to a profile one. Hence, this translation makes the recognition much easier since most reference databases include only frontal face samples. We made an End-to-End deep architecture that contains the GAN for translating profile samples and a ResNet-based classifier to identify the person from its synthesized frontal image. The experiments, which we conducted on an adequate dataset with respect to person-independent constraints between the training and testing, highlight significant improvement in the PIFR performance. Our contributions to the facial expression recognition task cover both static and

dynamic-based scenarios. The static-based FER framework relies on extracting textural and shape features from specific face landmarks that carry enough information to detect the dominant emotion. We proposed a new descriptor referred to as Orthogonal and Parallel-based Directions Generic Query Map Binary Patterns (OPD-GQMBP) to efficiently extract emotion-related textural features from 49 landmarks (regions of 32 by 32 pixels). These features are combined with shape ones computed by using Histogram of Oriented Gradients (HOG) descriptor on a binary mask representing the interpolation of the 49 landmarks. The classification is done through the SVM classifier. The achieved Person-Independent performance on five benchmarks with respect to Leave One Subject Out protocol demonstrated the effectiveness of the overall proposed framework against deep and handcrafted state-ofthe-art ones. On the other hand, dynamic FER contribution incorporates Long Term Short Memory (LSTM) deep network to encode the temporal information efficiently with a guiding attention map to focus on the emotion-related landmarks and guarantee the person-independent constraint. We considered four samples as inputs representing the evolution of the emotion to its peak. Each sample is encoded through a ResNet-based stream, and the four streams are joined by an LSTM block that predicts the dominant emotion. The experiments conducted on three datasets for dynamic FER showed that the proposed deep CNN-LSTM architecture outperforms the state-of-the-art.