博士論文要旨

論文題名：マルチモーダル深層学習を用いたカラー/深度画像によるジェスチャ認識と音声/表情によるうつ病検出

立命館大学大学院情報理工学研究科
情報理工学専攻博士課程後期課程

リュウ　カケイ
LIU Jiaqing

深層学習は、コンピュータビジョンや音声認識、自然言語処理などの様々な分野に活用され、その有効性が検証されてきた。そのほどんどはシングルモーダル学習に基づく手法である。一方、実応用において、様々なセンサーから得られるマルチモーダル情報が主であり，この考慮が必要である。そのため，マルチモーダル深層学習は，人工知能分野における重要な技術として，確立すべき課題である。現在のシングルモーダル深層学習に比べ、マルチモーダル深層学習には、マルチモーダルデータセットの構築、モダリティの表現、モダリティ間のアラインメント、モダリティ間の変換、マルチモダリティの融合、協働学習などの課題がある。本研究の目的は、効率的で正確なマルチモーダル深層学習法を開発し、医療・健康分野へ応用することである。主な成果を以下に示す。

(1) 三つの異なるマルチモーダルデータセットを構築した。一つ目は、マルチモーダルジェスチャデータセット(MaHG-RGBD: A Multi-angle view Hand Gesture RGB-D Dataset)である。２つの異なる角度に設置されたKinect V2から25種類のジェスチャーを撮像した。15人の成人被験者からそれぞれ100回ずつ、計75000ペア画像（カラーと深度のペア画像）により構成される。二つ目は、マルチモーダル姿勢データセット（Pose-RGBD）である。15種類の姿勢を6人の成人被験者から計13800ペア画像（カラーと深度のペア画像）をKinect V2から撮像した。三つ目は、うつ病傾向者のマルチモーダル情動データセット（MB-DD: multimodal behavioural dataset of depression）である。それぞれの被験者には情動データとうつ病に関するスクリーニング結果がある．情動データは、心理学専門家が設計したタスクに対する被験者の表情動画像、音声、歩行を対象としている。それらの情動データは、２台のビデオカメラと５台のマイクで記録した。

(2) 手術現場では、衛生状態を保ちながら執刀医自らが患者個々の医用画像に対する可視化操作が求められている。本研究では、手のジェスチャ認識によるタッチレス可視化法を提案し、システムを開発した。これまで、ジェスチャ認識法の改良を重ね，計三

つのバージョンシステムを開発した。バージョン１はKINECTで取得した手の深度画像に対して、HOG特徴量とSVMを用いたジェスチャ認識であった。バージョン２は認識速度を向上させるために、Microsoft社が提供するAPI関数を用いて、３つのジェスチャに限定した高速なタッチレス可視化システムであった。バージョン３は、バージョン1,２における課題の克服を検討し，システムの柔軟性と拡張性を向上させるために、Kinectセンサーから得られたカラー画像と深度画像の両方を用いたマルチモーダルジェスチャ認識法を提案した。従来のカラー画像または深度画像のみを用いたジェスチャ認識より高い認識精度を実現した。

(3) カラー画像のみを用いたジェスチャおよび姿勢の認識精度の向上を目指し、敵対的深層学習ネットワークを用いたカラー画像から深度画像への変換法を提案した。変換された擬似深度画像とカラー画像を用いた高精度なジェスチャ・姿勢認識法を開発した。提案法の有効性は、(1)の成果であるprivateデータセット（MaHG-RGBD, Pose-RGBD）と公開データセット（OUHANDS）を用いて検証した。ジェスチャ認識と姿勢認識の両方の実験においても有効であった。

(4) ストレスが多い現代社会では"うつ病”が大きな問題である．うつ病に対する早期発見と適切な治療により、大部分の改善が期待できることから，人工知能を用いた音声や表情などの特徴からうつ状態を検出する手法が期待されている。本研究では、被験者の音声と表情動画像を用いたうつ状態検出法（Multi-modal Adaptive Fusion Transformer Network）を提案した。主な貢献点は，①適応的マルチモダリティ特徴融合法を提案し、うつ状態検出に最も有効なモダリティ特徴を強調することによって高い検出精度の実現、②再帰型ニューラルネットワーク(RNN)の代わりにTransformerを用いたLong-range の時系列特徴を抽出による、検出精度の向上、③マルチタスク学習による検出精度の向上、である。提案法はstate-of-the-art精度を達成した。

Abstract of Doctoral Dissertation

Title：Multimodal Deep Learning Frameworks for Gesture Recognition in Color-Depth Images and Depression Detection in Audio-Visual Expressions

Doctoral Program in Advanced Information Science and Engineering
Graduate School of Information Science and Engineering
Ritsumeikan University

リュウ　カケイ
LIU Jiaqing

Deep learning has been successfully applied in many research field, such as computer vision, speech recognition and natural language processing. Most of them are focused on single modality. On the other hand, multimodal information is more useful for practical applications. Multimodal deep learning has gained a lot of attention and becomes an important issue in the field of artificial inetelligence. Compared to traditional single-modal deep learning, there are following challenges in multimodal deep learning such as: development of multimodal dataset; multimodal representation; multimodal alignment; multimodal translation and multimodal co-learning. The purpose of this research is to develop an efficient and accurate multimodal deep learning methods and apply them to the healthcare systems range from touchless medical image visualization for surgery to estimation of depression level using computer vision and deep learning. The main achievements of this research work are as follows.

(1) I developed three multimodal datasets for three different applications of multimodal deep learning. The first one is a multi-angle view hand gesture RGB-D dataset (MaHG-RGBD), which contains 75000 paired color-and-depth images of 15 subjects with 25 hand gestures obtained by two Kinect V2 sensors from different viewing directions. The second one is a human pose RGB-D dataset (Pose-RGBD), which contains 13800 paired color-and-depth images of 6 subjects with 15 postures obtained by Kinect V2 sensor. The third one is a multimodal behavioural dataset of depression (MB-DD), which comprises two components: the behavioural dataset and the screening survey results. The behavioural dataset contains dynamic expression facial images, speech and gait of depression subjects with different depression levels, which are recorded by two video cameras and five microphones.

(2) In medical surgery, surgery often faces the challenge of efficiently reviewing the patient's 3D anatomy model while maintaining a sterile field. I have proposed to use hand gesture recognition techniques to support, touchless visualization of 3D medical images in surgery. To achieve this, I have developed three version. The 1st version, I used HOG as feature and SVM as a classifier to recognize 9 kind of hand gestures from the depth images. In the 2nd version, the system uses a Kinect sensor to acquire three kind of hand gestures and track their hand movements. Based on

these states and their movements to visualize 3D hepatic anatomic models in real-time. In the 3rd version, I have proposed a multimodal deep learning method to perform gesture recognition using color and depth images. The multimodal system achieves more accurate and robust real-time gesture recognition compared with single-modal system.

(3) Image-based human posture recognition is a challenging problem due to many aspects such as cluttered background and posture self-occlusion. With the help of depth information, depth-based methods have better performance. However, depth cameras are not as widely used and not as affordable as color cameras. Therefore, I proposed a two-stage deep convolutional neural network (CNN) architecture for accurate color-based posture recognition. The first stage performs translation of color images to depth images, which is called as pseudo depth image. The second stage recognizes posture classes using both the color image and its pseudo depth image. The translation stage is based on a conditional generative adversarial network (cGAN). The proposed method was validated on two private datasets (i.e., Pose-RGBD, MaHG-RGBD) and one public dataset (i.e., OUHANDS). Experiments demonstrate that the proposed method achieves superior performance on both human pose and hand gesture recognition tasks.

(4) Depressive symptoms are a massive problem in this stressful modern society. Early screening of depressive symptoms helps to reduce the number and intensity of their depression episodes. Automatic detection of depressive symptoms from audio cues has gained increasing interest in the recent years. In order to achieve this, I have proposed a multimodal adaptive fusion transformer network for estimating the levels of depression. The proposed transformer-based network is utilized to extract long-term temporal context information from single-modal audio and visual data in our work. We also proposed an adaptive fusion method for adaptively fusing useful multimodal features. Furthermore, inspired by current multi-task learning works, we incorporate an auxiliary task (depression classification) to enhance the main task of depression level regression (estimation). The experimental results show that the proposed methods outperforming state-of-the-art methods.