Doctoral Dissertation

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

September 2021

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

LIU Jiaqing

Doctoral Dissertation Reviewed

by Ritsumeikan University

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

（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）

September 2021

2021 年 9 月

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

LIU Jiaqing
リュウ　カケイ

Supervisor: Professor Chen Yen-Wei
研究指導教員：　陳　延偉　教授

Abstract

Deep learning has been successfully applied in many research fields, such as computer vision, speech recognition and natural language processing. Most of them are focused on a 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 intelligence. Compared to traditional single-modal deep learning, there are the following challenges in multimodal deep learning such as: development of the 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 versions. The 1st version, I used HOG as the feature and SVM as a classifier to recognize 9 kinds of hand gestures from the depth images. In the 2nd version, the system uses a Kinect sensor to acquire three kinds 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 a 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 my work. I also proposed an adaptive fusion method for adaptively fusing useful multimodal features. Furthermore, inspired by current multi-task learning works, I 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.

博士論文要旨

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

一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。

Contents

Chapter 1

Introduction

Deep learning has been successfully applied in many research fields, such as computer vision, speech recognition and natural language processing. Most of them are focused on a 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 intelligence. Compared with traditional single-modal deep learning, there are the following challenges in multimodal deep learning: development of multimodal dataset; multimodal representation; multimodal alignment; multimodal translation; multimodal fusion and multimodal co-learning [1]. The purpose of this research is to develop efficient and accurate multimodal deep learning methods and apply them to 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 dissertation are shown in Figure 1.1. My research pipeline is based on build dataset, proposed original method for solving the multimodal challenge problems (indicated by black letters in the orange background) and apply to the healthcare system.

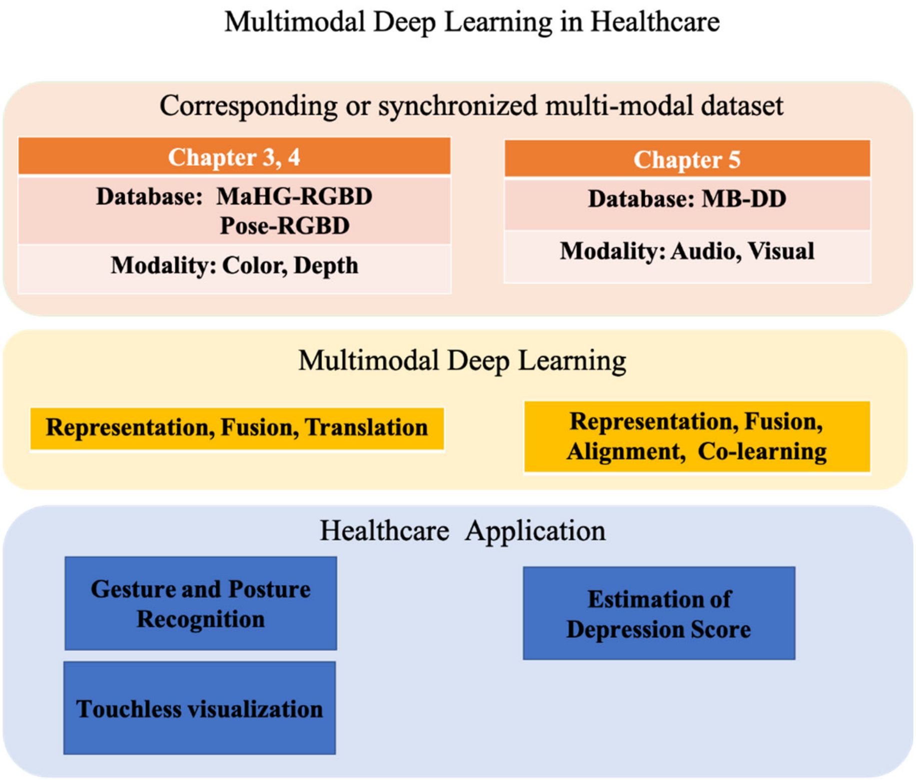

Multimodal Deep Learning in Healthcare
Corresponding or synchronized multi-modal dataset
Chapter 3, 4
Database: MaHG-RGBD
Pose-RGBD
Modality: Color, Depth
Chapter 5
Database: MB-DD
Modality: Audio, Visual
Multimodal Deep Learning
Representation, Fusion, Translation
Representation, Fusion, Alignment, Co-learning
Healthcare Application
Gesture and Posture Recognition
Touchless visualization
Estimation of Depression Score

Figure 1.1: Contributions on multimodal deep learning in healthcare. Black letters in the orange background the corresponding multimodal challenge tasks.

The main contributions of my research are 1) building three multimodal datasets: MaGH-RGBD hand gesture database [2], human pose RGB-D dataset (Pose-RGBD) [3] and multimodal behavioural dataset of depression (MB-DD) [4], 2) developing multimodal methods to solve the multimodal challenge problems: a) multimodal hand gesture recognition based on multimodal representation and fusion [5]; b) Translation of color image to depth image for accurate color-based posture recognition based on multimodal translation and fusion [3,6]; c) Extract and fuse the synchronized dynamic facial features associated with different emotion voice stimuli based on multimodal alignment and fusion [7]; d) an adaptive multitask fusion transformer network based on multimodal co-learning and fusion [8], 3) apply the multimodal methods to healthcare applications: touchless medical visualization system [9] and estimation of depression level [8].

In Chapter 2, I introduce the concept of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). I also present the basic idea about how to perform multi-modal fusion using early fusion and late fusion.

In Chapter 3, I propose a multi-angle view hand gesture RGB-D dataset (MaHG-RGBD), which contains 75,000 paired color-and-depth images of 15 subjects with 25 hand gestures obtained by two Kinect V2 sensors from different viewing directions. I proposed a multimodal deep learning method to perform the image recognition using a pair of color-and-depth images and apply them to touchless visualization of 3D medical images. The nine gestures that are associated with the high recognition accuracies were selected for the touchless visualization system. I further demonstrated that this technique facilitates touchless real-time visualization of hepatic anatomical models during surgery. This system is expected to ultimately lead to better patient outcomes by enhancing the ability to visualize medical images in 3D during surgery.

In Chapter 4, I demonstrated that the depth images provide higher recognition than the color image. Though the depth image is more useful and accurate for posture recognition than the color image, the depth cameras are not as widely used and not as affordable as color cameras. I proposed an RGB posture-recognition network based on a two-stage CNN architecture. To improve the recognition performance from color images, I generated an estimated depth posture image by a hybrid loss function incorporated in the generation module. The loss function captures the high-level features and recovers the sharp depth discontinuities. The proposed method was evaluated on the two datasets, including our novel dataset of color-depth pose images, and the public OUHANDS hand gesture dataset. The hybrid loss effectively and accurately generated depth posture images and the estimated depth image improved the accuracy of posture recognition.

In Chapter 5, I first introduced the basic methods of experimental design and data acquisition system of computer-aided depressive severity diagnosis. Second, I introduced the multimodal behavioral dataset of depression (MB-DD) [4], which comprises two components: the behaviours dataset and the screening survey results. The behavioural dataset contains dynamic expression facial images, speech, and gait of 102 subjects with different depression levels, which are recorded by two video cameras and five microphones. Third, I summarised the baseline behavioral features such as facial expressions and speech prosody and the baseline gate recurrent unit (GRU) network and a late fusion strategy to combine audio and visual modalities. Finally, I presented a multi-modal adaptive fusion transformer network for depression detection using multi-task representation learning with facial and acoustic features, which achieves the best results on the development set of the AVEC 2019 DDS dataset. By fusing the selected modalities, my proposed approach achieved a CCC score of 0.733 on the AVEC 2019 DDS dataset, outperforming the alternative methods investigated in this work.

In Chapter 6, I presented the conclusion of this dissertation.

Bibliography

1. Baltrušaitis, T., Ahuja, C., and Morency, L., "Multimodal Machine Learning: A Survey and Taxonomy," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019,
2. Liu, J.Q., Furusawa, K., Tsujinaga, S., Tateyama, T., Iwamoto, Y., Chen, Y.W., "MaHG-RGBD: A Multi-angle View Hand Gesture RGB-D Dataset for Deep Learning Based Gesture Recognition and Baseline Evaluations," Proc. of IEEE 37th International Conference on Consumer Electronics (IEEE ICCE2019), Las Vegas, USA, Jan. 11-13, 2019.
3. Liu, J.Q., Tsujinaga, S. Chai, S.R. Sun, H. Tateyam, T. Iwamoto, Y. Huang, X.Y. Lin, L.F. Chen, Y.W." Single Image Depth Map Estimation for Improving posture Recognition", IEEE Sensors, resubmitted
4. Liu, J.Q., Huang, Y., Huang X.Y., Xia, X.T., Niu X.X. and Chen, Y.W.,"Multimodal Behavioral Dataset of Depressive Symptoms in Chinese College Students–Preliminary Study," In: Chen YW., Zimmermann A., Howlett R., Jain L. (eds) Innovation in Medicine and Healthcare Systems, and Multimedia. Smart Innovation, Systems and Technologies, vol 145. Springer, Singapore, pp.179-190, 2019 Proc. of InMed2019, Malta, June 17-19, 2019.
5. Liu, J.Q., Furusawa, K., Tateyama, T., Iwamoto,Y,. and Chen Y.W., "An Improved Kinect-Based Real-Time Gesture Recognition Using Deep Convolutional Neural Networks for Touchless Visualization of Hepatic Anatomical Mode," *Journal of Image and Graphics,* Vol. 7, no. 2, pp. 45-49, 2019.
6. Liu, J.Q., Furusawa, K., Tateyama, T., Iwamoto,Y,. and Chen Y.W., "An Improved Hand Gesture Recognition with Two-Stage Convolutional Neural Networks Using a Hand Color Image and Its Pseudo-Depth Image," *Proc. of 2019 IEEE International Conference on Image Processing (IEEE ICIP 2019)*, Taibei, Taiwan, pp.375-379, Sep. 22-25, 2019.
7. Liu, J.Q., Huang, Y., Huang X.Y., Xia, X.T., Niu X.X. Lin, L.f., and Chen, Y.W., "Dynamic Facial Features in Positive-Emotional Speech for Identification of Depressive Tendencies" in Y.-W. Chen et al. (eds.), Innovation in Medicine and Healthcare, Smart Innovation, Systems and Technologies 192 (Proc. of InMed2020), pp.127-134 (2020).

8. Sun, H., Liu, J.Q., Chai,S.R., Qiu, Z.L., Lin, L.F., , Huang, X.Y., and Chen, Y.W., "Multi-modal Adaptive Fusion Transformer Network for the Estimation of Depression Level", *Sensors*, Vol.21, 4764, 2021. (https://doi.org/10.3390/s21144764). (co-first authors).
9. Liu, J.Q., Tateyama,T., Iwamoto,Y., and Chen,Y.W., "A Preliminary Study of Kinect-Based Real-Time Hand Gesture Interaction Systems for Touchless Visualizations of Hepatic Structures in Surgery," *Medical Imaging and Information Sciences*, Vol. 36, no. 3, pp. 128-135, 2019.

Chapter 2

Fundamentals of Deep Learning

2.1 Convolutional Neural Network

Convolutional Neural Networks (CNNs) are one of the most successful ideas in deep learning, generally including convolutional layers, pooling layers, and fully connected layers. They are made up of neurons that have learnable weights (*W*) and biases (*b*). They can be split into two parts: feature extraction part (convolutional layer and pooling layer) and classification part (fully connected layers). On the fully connected layer, they have a loss function. The image is first going through a series of hierarchically convolution layers, pooling layers for feature extraction. Thus, the extracted features are fed to the full connected layers for classification. The Figure 2.1 shows a typical architecture of CNN (AlexNet). It consists of five convolutional layers, three pooling layers (Max pooling), and two fully connected layers (FC).

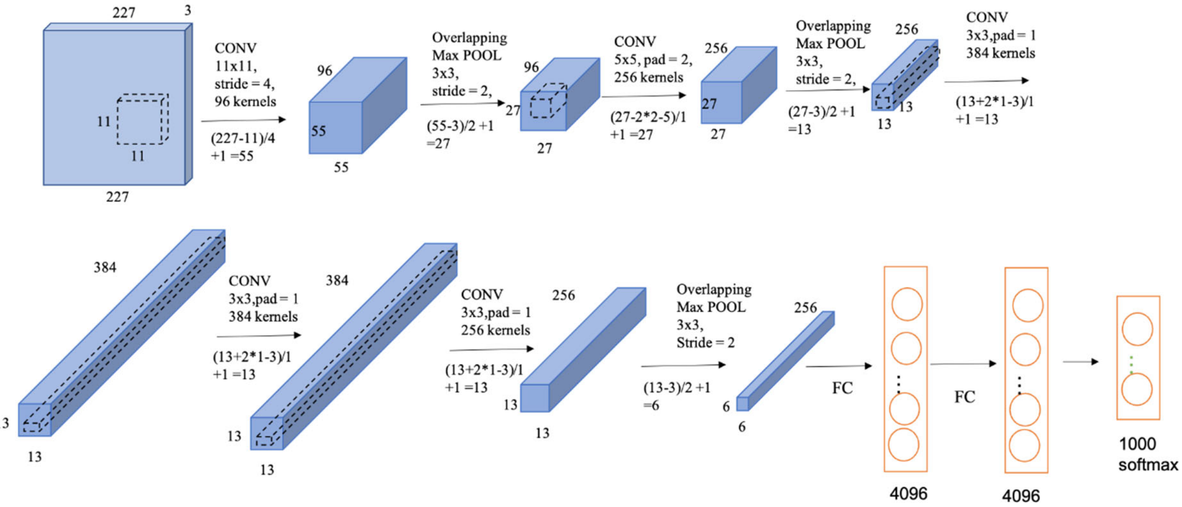


Figure 2.1: A typical convolutional Neural Network (AlexNet).

2.1.1 Convolutional Layer

The image that goes through the convolutional layers can be considered as feature extraction. The convolutional layer performs an operation called a "convolutional", which involves the multiplication of sets of weights with the input. The set of weights is called a filter or a kernel. The idea of applying convolutional operation is not new. Traditional filters are designed by experts based on their experience for a specific purpose. The traditional machine learning uses existing filters, such as Laplacian filter and Gabor filter, to extract features and the features are fed into the classifier for classification. Only the classifier is trained using training samples. The innovation of using the convolution operation in a neural network is that the weights of the filter are learned together with the FC layers (classifier) in a fashion of end-to-end. It means that we can automatically obtain optimum or specific filers to extract features for a given task Convolutional neural networks can learn multiple features in parallel for a given input. Different feature maps can extract different types of features. Each filter is called as a channel in the convolutional layer. In the first convolutional layer of AlexNet as shown in Figure 2.1, there 96 filers resulting in 96 feature maps, the output of the convolutional layer can be viewed as a volumetric image.

The size of the filter kernel is smaller than that of the input data, and an element-wise multiplication (dot production) is applied between a kernel size patch of the input and the kernel. The amount by which the filter shifts at each step is called stride. When the stride is 1, we move the filters one pixel at a time. Suppose the kernel size is *K* and the stride is *S*, the convolution operation can be expressed by Equation (2.1):

$$u(i,j) = \sum_{l=-(K-1)/2}^{(K-1)/2} \sum_{m=-(K-1)/2}^{(K-1)/2} w(l,m)x(i \times S + l, j \times S + m) \tag{2.1}$$

Where w, x, u are filter kernel, input, and output, respectively. Examples of convolution operation with strides 1 and 2 are shown in Figure 2.2.

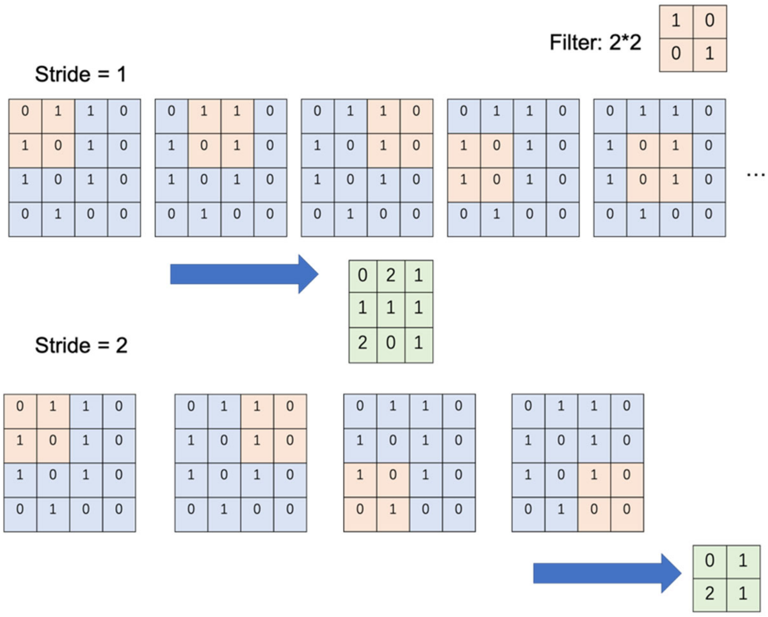
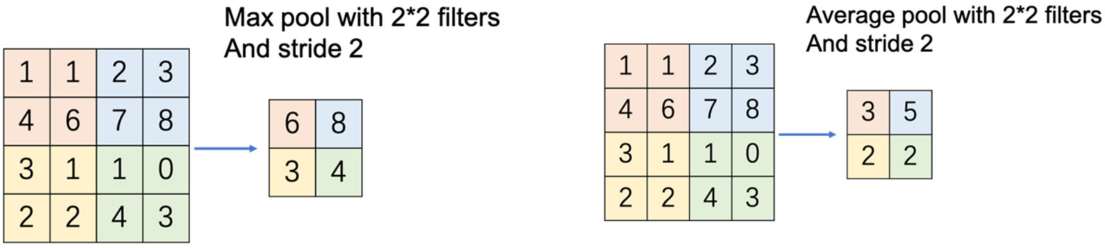


Figure 2.2: Convolution operations with stride of 1 and 2.

When we want to control the spatial size of the output volumes, we can use padding to surround the input with zeros. Suppose the size of the input image is *N,* the kernel size is *K*, and the stride and padding are *S* and *P*, respectively. The size of the output of the convolution is defined as (*N*-*K*+2*P*)/*S*+1. In Figure 2.2, the size of the output is (4-2+0)/1+1=3 for S=1 and (4-2+0)/2+1=2 for S=2, respectively. In the first convolutional layer of AlexNet (Figure 1), the size of the output is (227-11+2*0)/4+1=55. Since we have 96 channels (kernels), the size of the feature map (output volume) is 55 × 55 ×96.

2.1.2 Pooling Layer

In the CNNs, a pooling layer follows a convolution layer and has the same number of feature maps (channels) as the previous convolution layer. Each feature map in pooling layers subsampling on the feature map in the previous layer. Therefore, the pooling layer can effectively reduce the size of the feature maps and reduce the number of parameters in the last fully connected layer. The usage of the pooling layer can speed up the calculation and prevent over-fitting. The examples of max pooling and average pooling are shown in Figure 2.3 (a) and 2.3 (b), respectively.



(a) (b)

Figure 2.3: Examples of max pooling (a) and average pooling (b).

2.1.3 Fully Connected Layer

Fully connected layers connect every neuron in input layer to every neuron in output layer. It is in principle the same as the traditional multi-layer perceptron neural network. The fully connected layers are often used in the classification task, which is the final part of the CNN, it takes the output of formal layers as inputs, and maps them into the targets of the classification task as output.

In this way, the CNN transforms the original pixel values from the original image layer by layer to the final classification results.

2.2 Recurrent Neural Networks

In feedforward neural networks, which are discussed in section 2.1, data is processed only the way from input to output. In contrast, Recurrent neural networks (RNN) are primarily used to process time-series data. RNNs include a feedback loop that sends the output of processed information back as an input at the next time step in the sequence. The basic idea of RNN is shown in Figure 2.4. We can process a sequence of vector *x* by applying a recurrence formula at every time step. The same function and the same set of parameters are used at every time step. The hidden state update process can summary as Equation (2.2):

$$h_t = f_W(h_{t-1}, x_t) \tag{2.2}$$

where x_t donates the input vector at *t* time step h_t represent the new state, h_{t-1} represent the old state, f_W corresponds to some non-linear transformation such as tanh, ReLU with parameters *W*.

The advantages of RNN are: 1) can process any length input, 2) same weights applied to every timestep. Model size doesn't increase for longer input. On the other side, the disadvantages of RNN are: 1) Recurrent computation is slow, 2) it difficult to access information from many steps back, which means it is difficult to extract long-term temporal context information from long sequences.

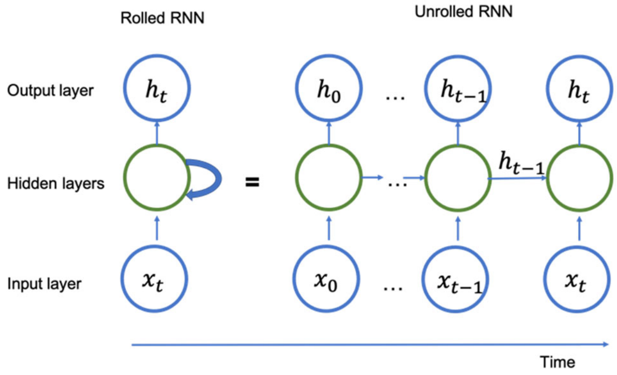


Figure 2.4: Basic structure of Recurrent Neural Network.

2.2.1 Long Short-Term Memory

The design of Long short-term memory (LSTM) [1] is inspired by logic gate of a computer. Basic elements of Long short-term memory include an input gate to control activations for the memory cell, a forget gate to drop useless information of the past cell status, and an output gate to control the output activations for the ultimate state (Figure 2.5).

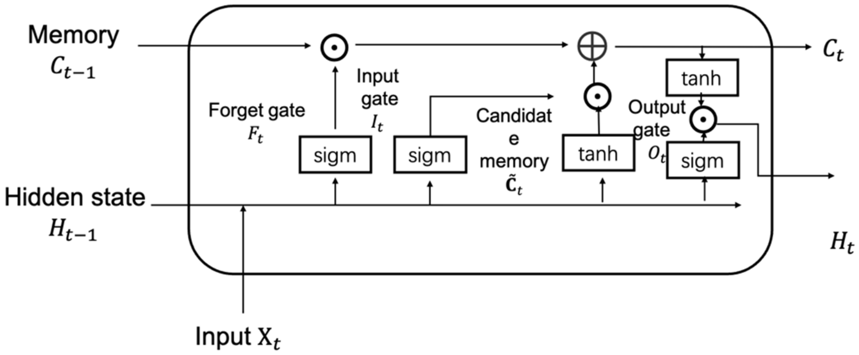


Figure 2.5: The unfolded chain structure of LSTM in time sequence

The update of LSTM units at time-step t can be described as the Equation (2.3). where $\mathbf{I}_t$ is the input gate, $\mathbf{F}_t$ is the forget gate, $\mathbf{O}_t$is the output gate, σ is the logistic sigmoid function, $\mathbf{W}_{xi}, \mathbf{W}_{xf}, \mathbf{W}_{xo}$ *and* $\mathbf{W}_{hi}$, $\mathbf{W}_{hf}, \mathbf{W}_{ho}$ are weight parameter, $\mathbf{b}_i$, $\mathbf{b}_f, \mathbf{b}_o$ are biases. The input is $\mathbf{X}_t$ and the hidden state of the pervious time step is $\mathbf{H}_{t-1}$.

$$\begin{aligned} \mathbf{I}_t &= \sigma(\mathbf{X}_t\mathbf{W}_{xi} + \mathbf{H}_{t-1}\mathbf{W}_{hi} + \mathbf{b}_i) \\ \mathbf{F}_t &= \sigma\left(\mathbf{X}_t\mathbf{W}_{xf} + \mathbf{H}_{t-1}\mathbf{W}_{hf} + \mathbf{b}_f\right) \\ \mathbf{O}_t &= \sigma(\mathbf{X}_t\mathbf{W}_{xo} + \mathbf{H}_{t-1}\mathbf{W}_{ho} + \mathbf{b}_o) \end{aligned} \tag{2.3}$$

The LSTM architecture has the candidate memory cell $\tilde{\mathbf{C}}_t$. Its calculation is similar to the three gates described above but using a tanh function as the activation function. $\tilde{\mathbf{C}}_t$ can summary as Equation (2.4):

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{X}_t\mathbf{W}_{xc} + \mathbf{H}_{t-1}\mathbf{W}_{hc} + \mathbf{b}_c) \tag{2.4}$$

where $\mathbf{W}_{xc}, \mathbf{W}_{hc}$ are weight parameters and $\mathbf{b}_c$ is a bias parameter.

In LSTM, we have the input gate $\mathbf{I}_t$ controls how much of the new data into account via $\tilde{\mathbf{C}}_t$ and the forget gate $\mathbf{F}_t$ addresses how much of the old memory cell content $\mathbf{C}_{t-1}$. The memory cell $\mathbf{C}_t$ can summary as Equation (2.5):

$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t \tag{2.5}$$

Another component of LSTM is hidden state. he hidden state $\mathbf{H}_t$ can summary as Equation (2.6):

$$\mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t) \tag{2.6}$$

where $\odot$ is the elementwise product operator. When the output gate $\mathbf{O}_t$ close to 0, the network retains all the information only within the memory cell $\mathbf{C}_t$. When the gate $\mathbf{O}_t$ close to 1, the networks pass all memory information through to the predictor.

2.2.2 Gated Recurrent Unit

The gated recurrent unit (GRU) [2] is the newer RNN variant that make it much better capturing long range connection and solve with the vanishing gradient problems. GRU is got rid of the cell state and used the hidden state to transfer information. It also has two gates. a reset gate and update gate. The GRU Cell and its gate is shown in Figure 2. 6. The reset gate is used to decide how much past information to forget. The update gate decides what information to throw and what new information to add. It is similar to the forget and input gate of an LSTM. Then, the reset gate $\mathbf{R}_t \in \mathbb{R}^{n\times h}$and update gate $\mathbf{Z}_t \in \mathbb{R}^{n\times h}$ are computed as follows:

$$\begin{aligned}\mathbf{R}_t &= \sigma(\mathbf{X}_t\mathbf{W}_{xr} + \mathbf{H}_{t-1}\mathbf{W}_{hr} + \mathbf{b}_r),\\ \mathbf{Z}_t &= \sigma(\mathbf{X}_t\mathbf{W}_{xz} + \mathbf{H}_{t-1}\mathbf{W}_{hz} + \mathbf{b}_z),\end{aligned} \tag{2.7}$$

where $\mathbf{H}_{t-1}$ is the hidden state of pervious time step. $\mathbf{W}_{xr}, \mathbf{W}_{xz,}\mathbf{W}_{hr}, \mathbf{W}_{hz}$ are weight parameters and $\mathbf{b}_r, \mathbf{b}_z$ are biases.

The candidate hidden state $\tilde{\mathbf{H}}_t$ is calculated as follows:

$$\tilde{\mathbf{H}}_t = \tanh(\mathbf{X}_t\mathbf{W}_{xh} + (\mathbf{R}_t \odot \mathbf{H}_{t-1})\mathbf{W}_{hh} + \mathbf{b}_h) \tag{2.8}$$

where $\mathbf{W}_{xh}, \mathbf{W}_{hh}$ are weight parameters and $\mathbf{b}_h$is bias. $\odot$ is the elementwise product operator. Compare with RNN, the influence of the pervious states can be reduced with the elementwise multiplication of $\mathbf{R}_t$ and $\mathbf{H}_{t-1}$.

The final update Equation (2.9) for the GRU is calculated as follows:

$$\mathbf{H}_t = (1 - \mathbf{Z}_t) \odot \mathbf{H}_{t-1} + \mathbf{Z}_t \odot \tilde{\mathbf{H}}_t \tag{2.9}$$

If $\mathbf{Z}_t$ is close to 1, $\mathbf{H}_t$ is closed to the candidate hidden state $\tilde{\mathbf{H}}_t$. If $\mathbf{Z}_t$ is close to 0, the GRU retain the old state $\mathbf{H}_{t-1}$. The GRU can cope with the vanishing gradient problem in RNNs and better capture long-rang sequence. Illustrates of GRU cell and its gates are shown in Figure 2.6.

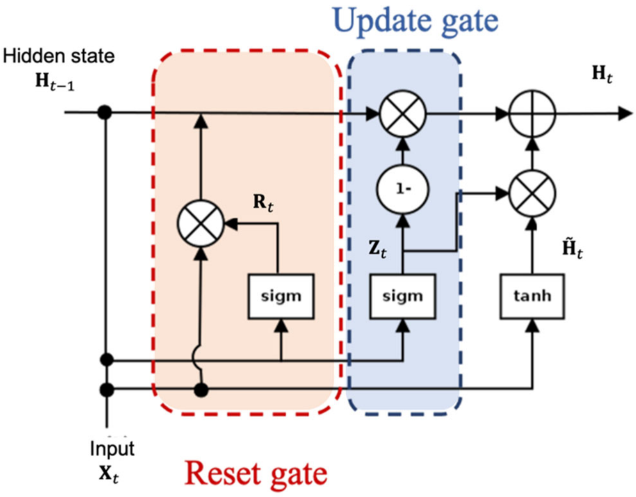


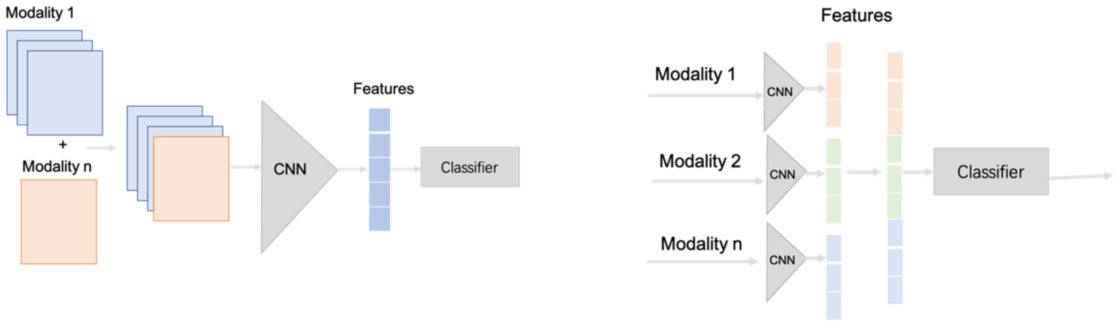
Figure 2.6: GRU cell and its gates. The orange part is reset gate, while the light blue part is update gate.

2.3 Multimodal fusion

Different sensors can provide different information about the same context. Multimodal fusion is the technology to join the relevant information from the different modalities that leads to accurate prediction over using only one modality [6,7]. The respective approaches can be broadly categorized as early fusion, late fusion, and intermediate fusion, depending on the position of the fusion.

2.3.1 Early fusion

Early fusion combines the different modalities before attempting to classify the content. There are two types of early fusion. One is using raw data, for example multi-modal images are used as multi-channel images (each modality is used as an input of channel). Another one is concatenating multimodal feature vectors into a joint representation and fed into a classifier for classification. The two types of early fusion architectures are shown in Figure 2.7 (a) and (b), respectively. Note that the first type is a single-stream architecture and is easy for implementation, but it can only be used for multi-modal data with same dimension. On the other hand, the second type is a multi-stream architecture, which may take more computation cost, but can be used for any multi-modal data even with different dimension such as fusion of audio and visual data.



(a) (b)

Figure 2.7: Two types of early fusion architecture. (a) using raw data; (b) concatenating multimodal feature vectors.

2.3.2 Late fusion

In late fusion, each modality is processed in a separate unimodal CNN stream and the scores (results) of each modality are fused into a final decision using a simple mechanism such as voting and averaging [8-11]. Late fusion is the simplest and most used fusion method. Late fusion has a major drawback which is the very limited potential for the exploitation of the cross-correlation between the different unimodal data. Figure 2.8. shows the basic late fusion architecture.

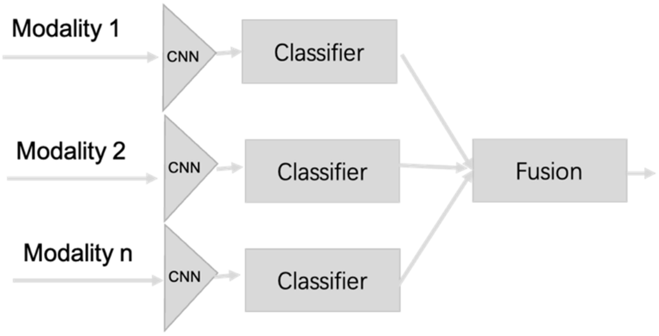


Figure 2.8: Late fusion architecture.

Bibliography

1. LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P.: Gradient-based learning applied to document recognition, Proceedings of the IEEE, Vol. 86, No. 11, pp. 2278-2324 (1998)
2. Hochreiter,S., Schmidhuber, J.,; Long Short-Term Memory. Neural Comput 1997; 9 (8): 1735–1780.
3. Cho, K., Merrienboer, B.V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation"
4. Lahat, D., Adali, T., and Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects, Proceedings of the IEEE, Vol. 103, No. 9, pp. 1449-1477 (2015)
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. arXiv preprint arXiv:1706.03762 2017.
6. Andrej K., George T., Sanketh S., Thomas L., Rahul S., and Li F.F., "Large-scale video classification with convolutional neural networks." In CVPR, 2014.
7. Yang, X.D., Molchanov, P., and Kautz. J., "Multilayer and multimodal fusion of deep neural networks for video classification." In International conference on multimedia, 2016.
8. Liu M., and Yuan. J., "Recognizing human actions as the evolution of pose estimation maps". In CVPR, 2018.
9. Abavisani, M., Joze, H.R.V., and Patel.V.M., "Improving the performance of unimodal dynamic handgesture recognition with multimodal training. "In CVPR, 2019.
10. Katsaggelos, A.K., Bahaadini, S., and Molina. R., "Audiovisual fusion: Challenges and new approaches". Proceedings of the IEEE, 103(9), 2015.
11. Emilie, M, Amaury H, and Stephane A," Majority vote of diverse classifiers for late fusion." In Structural, Syntactic, and Statistical Pattern Recognition, pages 153–162, Berlin, Heidelberg, 2014. Springer.

Chapter 3

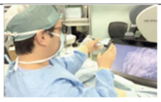
Multimodal Deep Learning for Accurate Gesture Recognition Using Color and Depth Images for Touchless Visualization of 3D Medical Image

3.1 Introduction

Understanding the patient's anatomic structure is essential for successful surgery [1-2]. Though the visualization of the reconstructed anatomic model on computers can provide detailed and useful anatomic information for surgery, the surgeon usually needs to use some contacting devices such as a mouse, keyboard, or touch panel to display the medical images during the surgical operation. After operating the visualization device, re-sterilization is necessary to maintain hygiene, which is an inefficient and un-effective process for surgery. Touchless technology is an attractive and potential solution to address the above problems. How to develop a real-time and accurate hand gesture interaction system is becoming the main challenge for touchless interaction in sterile environments for surgeons.

A lot of visualization and virtual reality techniques have been proposed for surgical navigation and surgical support. Sugimoto et al. [9] proposed a spatial navigation system for medical information by interactively superimposing a 3D hologram and 3D printing technology. The limitation is that users need to hold a pen connected to the Z-space system (VR display) as an interactor. After operating the visualization device, re-sterilization is necessary to maintain hygiene, which is an inefficient and un-effective process for surgery.

Table 3.1: Summary of state-of-the-art devices using in the area of touchless interaction in the operating room.

Device		Evaluation results and problems
3D printer [2]		It is possible to intuitively confirm the anatomic structure from a ee angle by holding the printed model in hand, but it lacks flexible sualization such as zoom in/out, selectivity of the specific vessel compared with the visualization of 3D models on computers.
Z-space [9]		Sugimoto et al. proposed a spatial navigation system for medical by superimposing 3D hologram and 3D printing technology. The limitation is that users need to hold a pen connected to the system as an interactor.
HoloLens [13]		HoloLens can detect hand gestures and realize the touchless visualization. The problem for HoloLens-based touchless visualization system is that surgeons must wear the HoloLens during the operation, which is not practical and will limit the operational performance.

Touchless technology is an attractive and potential solution to address the above problems. In 2010, Microsoft released a low-cost RGB-D camera, called Kinect. Kinect can provide both color image and depth image, and it can detect the human actions and human skeleton without any markers. Since it can be used for accurate gesture recognition, Kinect is considered as an ideal solution for touchless interactions. Several touchless interaction systems based on Kinect have been proposed for the visualization of medical images in the surgical operating room. Gallo et al. developed a controller-free exploration of medical image data [3]. Yoshimitsu et al. developed a system called “OPECT” for the visualization of 2-D slice images in brain surgery [4]. Roppurt et al. developed a touchless gesture user interface for interactive image visualization in urological surgery [5]. However, these systems still have some limitations: need two hands for interaction, slow responding time, lack of flexibility of interaction. So the purpose of this research is to develop a high accuracy, real-time, and flexible interactive touchless medical visualizing system using multimodal gesture recognition.

We have developed several versions to achieve our final goals. In the first version [6], I used HOG as feature extraction and SVM as classifier to recognize nine kind of hand gesture from the depth images only, the mean recognition accuracy is found to be 87.5% totally in 8 fps. The system could not achieve real-time recognition. In our second version (Section 3.3), I just recognized three kinds of hand states and their movements by using the API of the Kinect without processing of feature extraction and classification to realize a real-time interaction. They could not be able to realize complicate interaction and lack flexibility of interactions.

In the third version (Section 3.4), I built a novel Multiview RGB-D dataset, namely MaHG–RGBD. Moreover, I performed a recognition experiment to recognize depth images using deep learning. MaHG–RGBD comprises 25 classes of gestures, proposing 9 classes of gestures with high recognition accuracy. However, the average recognition accuracy of 9 and 25 classes was 96.51% and 91.87%, respectively. These showed that some gestures were difficult to recognize using only a depth image. I have proposed a two-stream multi-modal deep learning and to fully utilize the depth and color information. Thus, the proposed system outperforms my previous system from the viewpoint of recognition accuracy, rapidity, and flexibility.

The remainder of this Chapter is as follows. Section 3.2 introduces the touchless visualization system. Section 3.3 describes a real-time interaction based on 3 hand states combined with movements. Section 3.4 describes my originally proposed multi-angle view hand gesture RGB-D dataset for the deep learning-based gesture. In Section 3.4, I focus on the multimodal deep learning method to recognize the hand gestures. I conclude in Section 3.5 with a summary of my main findings and identify directions for future research.

3.2 Visualization system

3.2.1 System configuration

I designed a visualization module and an interaction module respectively to raise the usability and freedom of their adaptation to the surgery environment. The two modules communicate with each other through a socket. The hardware for the visualization module consists of a server PC with visualization software and a 3D display or a screen with a projector. As a demonstration system, I use an L-shaped stereoscopic display with two projectors connected to a server PC to display the 3D models. It should be noted that the L-shaped stereoscopic display with two projectors probably too big and not suitable for use in a surgical room. I use it just for the demonstration of 3D visualization. In real clinical applications, we may use a glass-free 3D display such as magnetic 3D [17] instead of the L-shaped stereoscopic display with two projectors. The hardware for the interaction module is a Microsoft Kinect connected with a smart PC, which is used to capture and recognize hand gestures. Our demonstration system is shown in Figure 3.1.

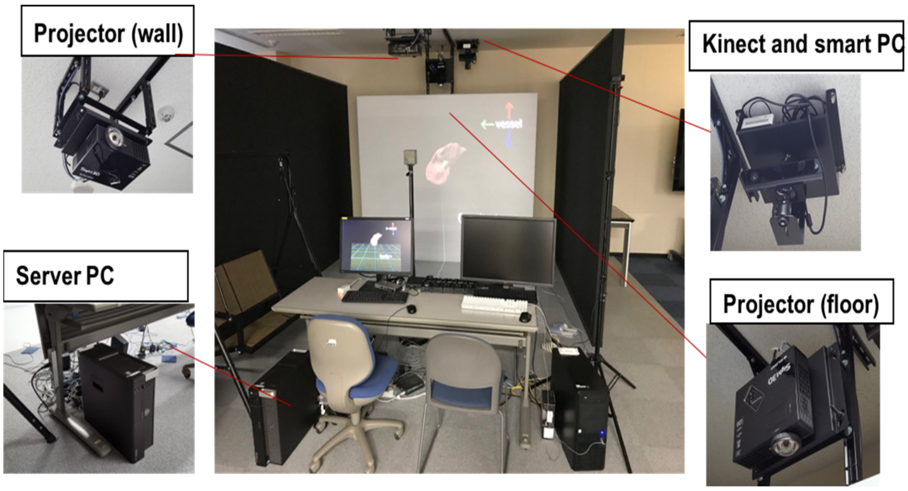


Figure 3.1: Our demonstration system.

The diagram in Figure 3.2 summarizes our system architecture that includes two modules: the interaction module and the visualization module. When the Kinect sensor detects that user's gesture becomes available state (i.e., user's right hand is above the waist for 45cm), it performs real-time hand gesture recognition and records hand's 3D location (in the Kinect's coordinate frame), the hand state and its movement are processed by command module and send to visualization modules through a socket, and finally, the visualization module responds to the command and performs the corresponding operation like rotation, opacity adjustment, zoom in/out, fusion and selection of vessels.

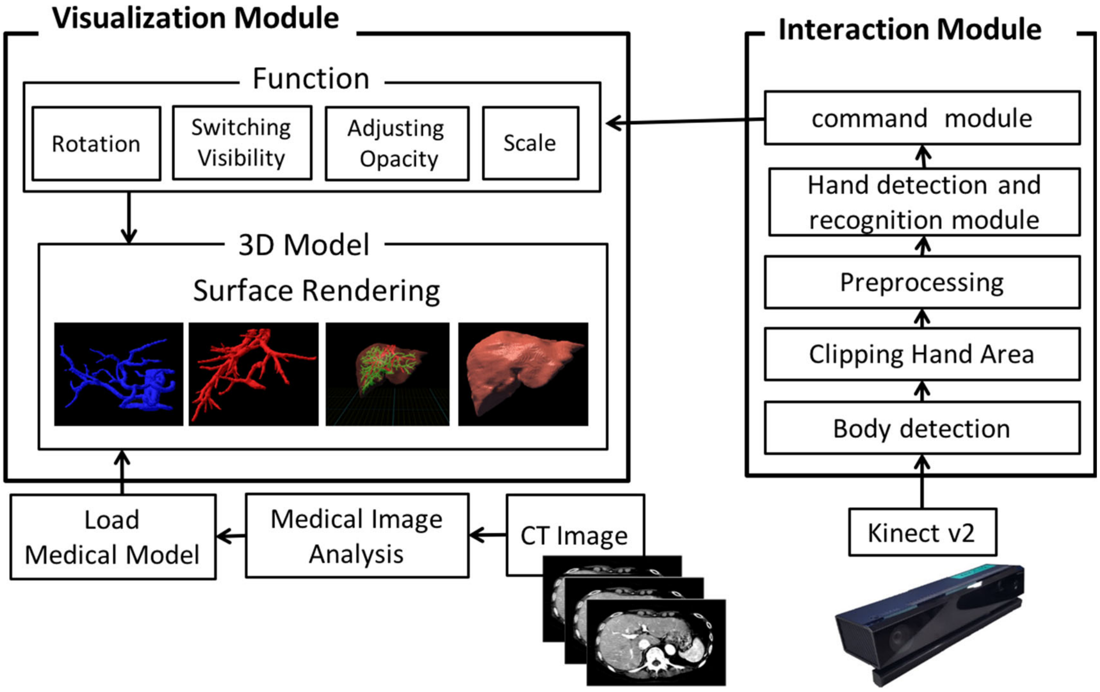
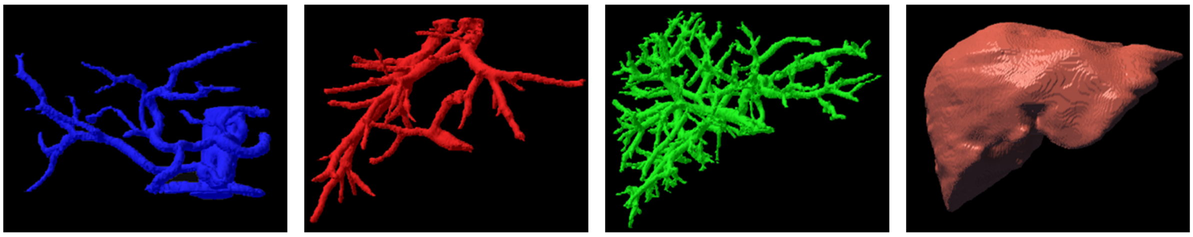


Figure 3.2: Diagram of the proposed system.

3.2.2 Visualization module

In the visualization module, surface models of hepatic structure including hepatic artery, hepatic portal vein, hepatic vein, and liver parenchyma (Figure 3.3) are generated by converting each corresponding volume data to a triangulated mesh surface using marching cube algorithms. Each volume of data is segmented semi-automatically from CT images under the guidance of a physician [1, 2].

Compared with the traditional slice-by-slice visualization and review techniques, the surgeon can easily recognize the liver geometry, its vessels structures, and locations during the surgery with the 3D surface rendering of hepatic structures as shown in Figure 3. Please refer to [1, 2] for detailed information about CT data and segmented liver and vessel data. The system has four visualization modes: rotation, zoom in/out, adjustment of opacity, fusion, and selection of vessels.



(a) Hepatic artery (b) Hepatic portal vein (c) Hepatic vein (d) Liver

Figure 3.3: Visualization of liver and its vessels.

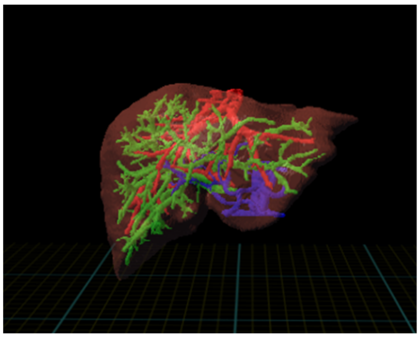


Figure 3.4: Visualization of fused liver and its vessel structure.

3.2.3 Interaction module

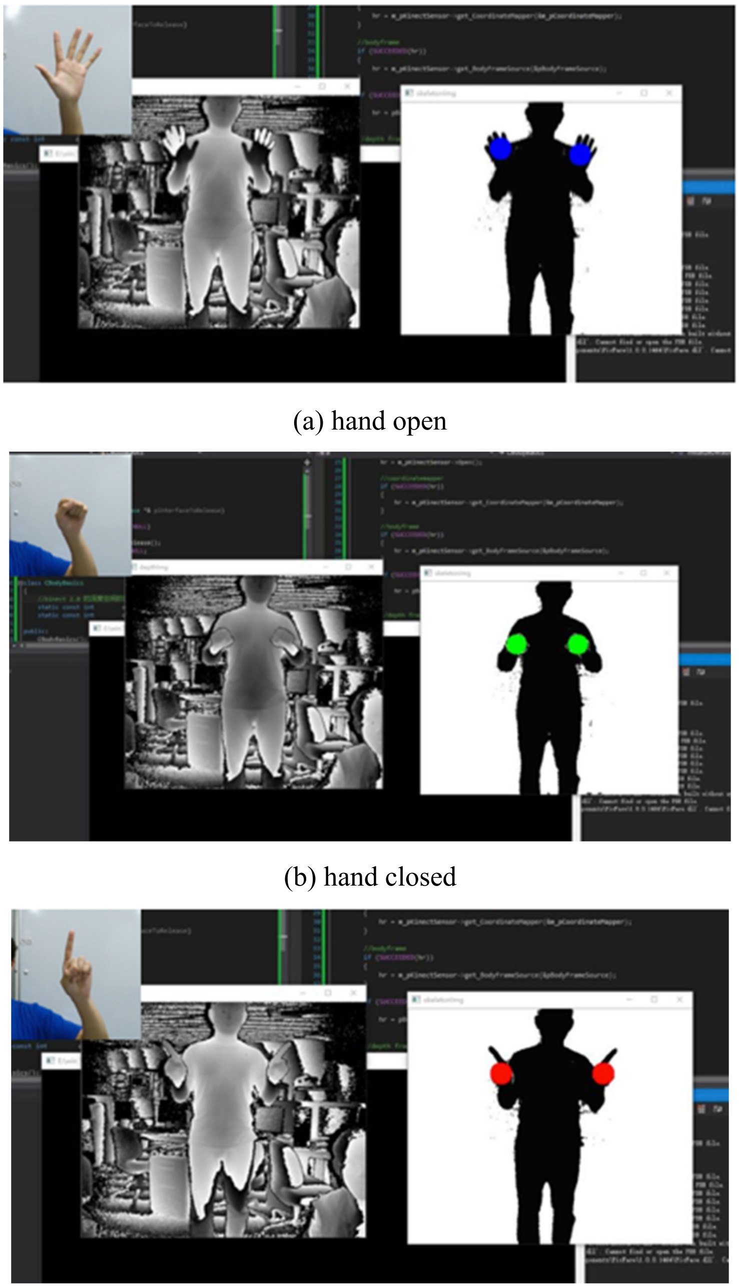
The interaction module is the main contribution of this research. In the first version [6], I used the histogram of oriented gradients features and a support vector machine (SVM) classifier. The method consisted of two processes: feature extraction and classification. Though machine learning-based methods achieved high recognition accuracy, they could not achieve real-time recognition. In the second version [22], I recognized three kinds of hand states and their movements by using the API of the Kinect without processing feature extraction and classification to realize a real-time interaction. Though the first proposed system has limitations: it is not able to realize complicated interactions and lacks the flexibility of interactions, the proposed interactions are enough for touchless visualization control. The originality and novelty of this preliminary study is that I proposed an easy and fast framework to solve this task without doing gesture recognition by ourselves. In the third proposed system [23, 24], I build a new dataset, which is recorded with 15 participants performing all twenty-five hand gestures. I use a multimodal deep learning technique to recognize hand gestures using the depth learning network that adds color information to depth information. A rapidly responding and flexible Kinect-based touchless visualization system has been realized.

3.3 Kinect-Based Real-time Hand Gesture Interaction Systems for Touchless Visualization of 3D Medical Image

3.3.1 Proposed method

In this section, I focused on the second version. To realize a real-time hand gesture interaction and visualization, I combined three hand states (open, close, and lasso), which are automatically detected by Kinect, with their hand movements to control the visualization in the new system. The hand gestures (hand state + hand movement) for controlling visualization mode are summarized in Table 3.2. Detailed information about visualization is described in Sec.3.2.3.

Since the Kinect V2 supports 3 types of hand states: open, closed, lasso (lasso is defined by closing the hand and extending the index finger), which are shown in Figure.3.5(a), (b) and (c), respectively. I can accurately recognize the hand state at high speeds. In addition to three types of hand states, I also use the movement of hand joints for hand gesture recognition. By using the function of skeleton tracking features in Kinect for Windows Software Development Kit (SDK) 2.0 [7], I can easily detect 31 landmarks of the skeletal human body (a machine-learning-based algorithm automatically interprets each pixel as belonging to the background or to one of the 31 parts in person's body has been subdivided [8]. This information is then used to calculate the position of the skeleton). I use HandRight and HandLeft joint points to detect hand movements. The hand movement detection algorithm is shown in Algorithm 3.1.



(a) hand open

(b) hand closed

(c) hand lasso

Figure. 3.5: Three types of hand states used in the system. (a) open, (b) closed, (c) lasso.

Algorithm 3.1 Hand movement detection algorithm

```
System starts:
visualization server loads image data and visualizes liver and its vessel structure
While server is listening to client (interaction module) do
  if user's hand is above waist for 45cm then
    user's gesture become available state
    else if hand from other states to open
        client records hand's location as a new benchmark
    else if hand state is close
        compare right hand's location this frame with benchmark frame hand location
        if difference above threshold, then
          send corresponding movement (both direction and distance) to server
        else
          continue
    else
      responding to other gestures
  else
    send neutral command
```

The visual information provided by the system is based on surgeons' advice and requirements. Especially the hepatic vessel structure is one of the most important visual information for hepatic surgery. I design simple and fast interactions for touchless operations. The interactions are designed based on the workflow of hepatic surgery. The interaction tasks that I observed most frequently during operation. I classified interaction steps during each procedure, such as rotation of 3D liver model to check it from different angles, adjustment opacity to check the structure of vessels inside, fusion and selection of vessels to confirm the positional relationship between them or zooming to analyse details in the models. The hand gestures (hand state + hand movement) for controlling visualization mode are summarized in Table 3.2. Right hand open and left hand open are idle states that ensure the system's ability to detect the next frame movement accurately.

There are two operations that use motion: (1) rotation and (2) fusion and selection of vessels. For rotation, if the direction of motion is around only one particular axis (x or y), the model will rotate around that axis. The angle of rotation is proportional to the movement's distance ($\Delta \mathrm{d} = x - x_0$ or $(y - y_0)$), which is represented by Equation (3.1):

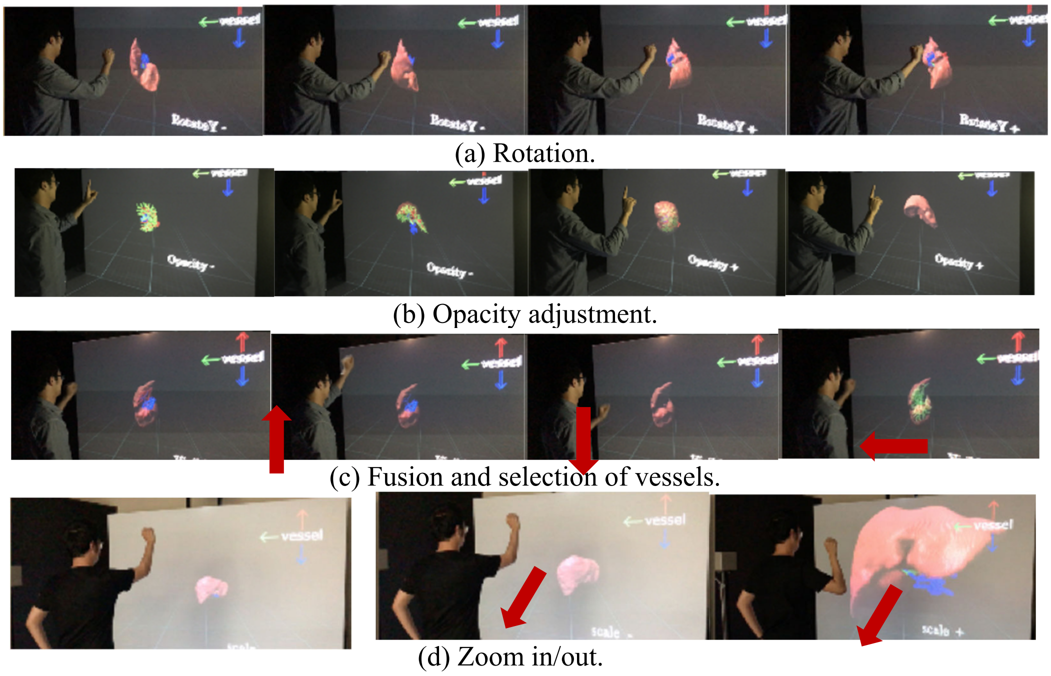
$$angle = \begin{cases} 0 & (|\Delta \mathrm{d}| < |\mathrm{D}|) \\ (\Delta \mathrm{d} - \mathrm{D}) * 10 \quad (\mathrm{degree}) & (|\Delta \mathrm{d}| > |\mathrm{D}|) \end{cases} \tag{3.1}$$

where D is the threshold set to ± 10 cm. So, the range of motion for rotation is +10 to +40 cm or −10 to −40 cm from the initial position x_0 (or y_0) in the predefined recognition area (−40 to

+40 cm). For the fusion and selection of vessels, the movement threshold is set to 20 cm. The range of motion is 20 to 40 cm.

Table 3.2: Visualization mode controlled by gestures.

Gesture	Interactions
right hand open	idle state for rotation
left hand open	idle state for selection of vessels
right hand closed and move 4-dimensional	rotation of models along corresponding direction
left hand closed and move left, up, down	fusion and selection of vessels
right hand in lasso state	opacity up
left hand in lasso state	opacity down
right hand pull back in closed state	zoom in
right hand push forward in closed state	zoom out



(a) Rotation.

(b) Opacity adjustment.

(c) Fusion and selection of vessels.

(d) Zoom in/out.

Figure 3.6: Examples of 4 visualization modes controlled by specific gestures.

3.3.2 Experimental setup

I conducted both single- and multi-user experiments, whose setups are shown in Figure. 3.7 and 3.8, respectively. In the single-user experiments, the distance between the Kinect sensor, which was 2.5 m in height and tilted 45° horizontally, and the user was 2.0 m. At most, six people could be detected at an effective depth range of 0.5 to 4.5 m. The Kinect field of view was 70.6° and 60° in the horizontal and vertical directions, respectively [18]. There were three users (A, B, and C) in the multi-user experiments, as shown in Figure 3.8. User A was the surgeon, whereas Users B and C were medical workers, who stood at 0.5 m to the right of and 0.5 m behind the surgeon, respectively.

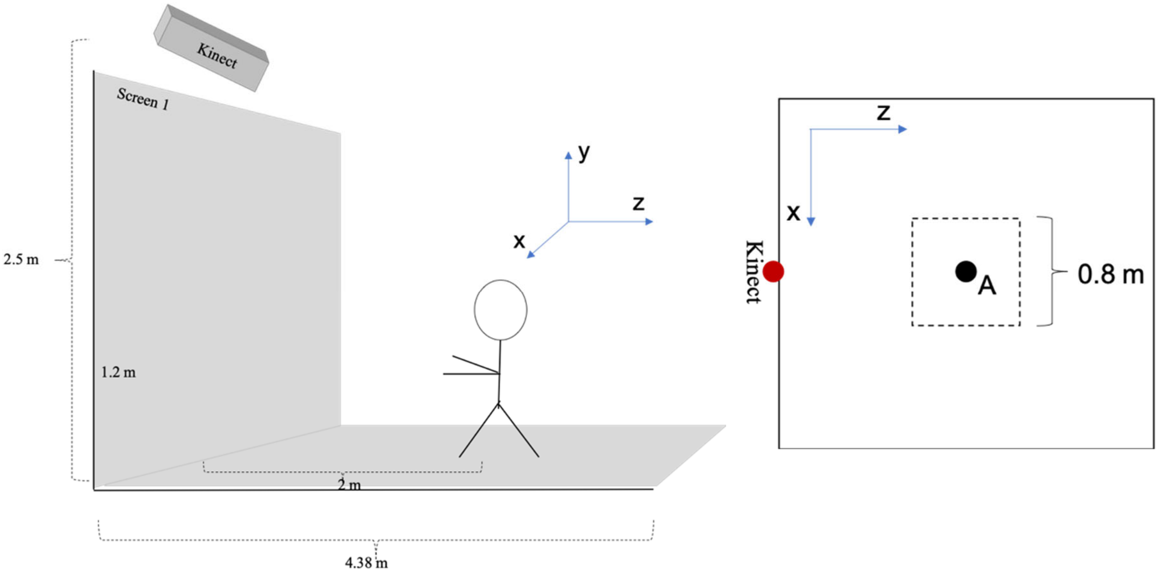


Figure. 3.7: (a) Experimental setup for a single user. (b) The dashed rectangle shows the recognition area in the x–z plane.

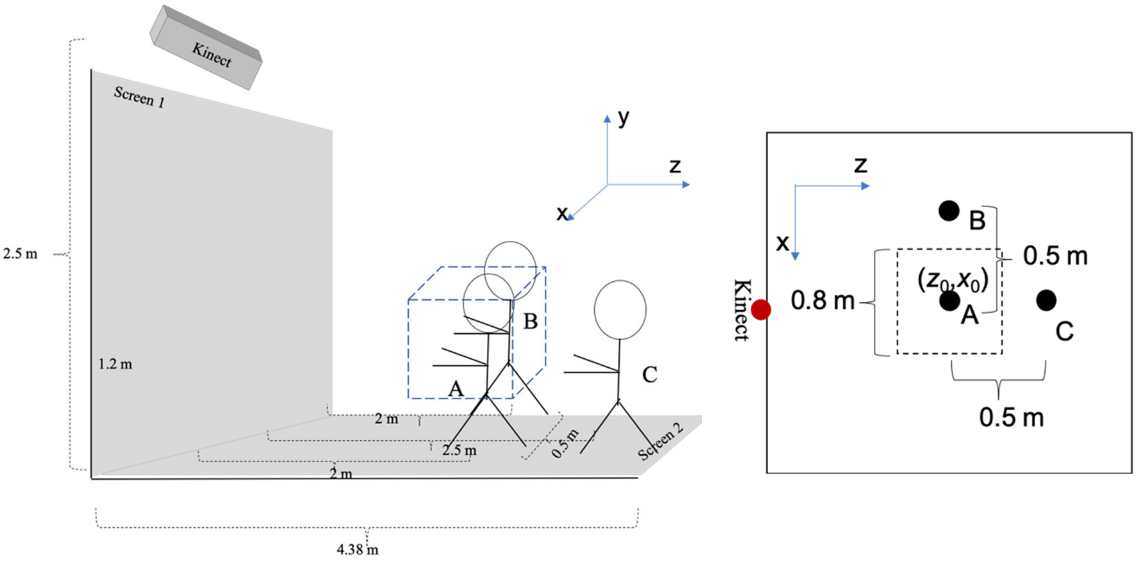


Figure. 3.8: (a) Experimental setup for multiple users. (b) The dashed rectangle shows the recognition area in the x–z plane.

3.3.3 Results

The participant was asked to make and repeat a different gesture (open, closed, or lasso) every 4 s. To simplify the problem, I used class ID to represent each gesture. The capture frame rate of Kinect was 30 fps. The results of the recognition by my previous and proposed systems [6] are shown in Figure. 3.9. The recognition rate of my first version system was 87.5% at 8 fps. In order to increase the recognition rate, I recognized the gesture within a temporal sliding window (rather than frame by frame). The most voted gesture was the final recognition result. I used a sliding window with 10 sequential frames, which corresponds to 0.8 fps. The recognition rate was improved by 100% at 0.8 fps. In contrast, my second version system achieved a recognition rate of 100%, even at 30 fps (real-time). By predefining the recognition area, the recognition accuracy for multiple users was the same as that for a single user. The measurements of both experiments were taken by a computer equipped with an Intel Core-i7 processor, 16GB of RAM, and an integrated graphics processor.

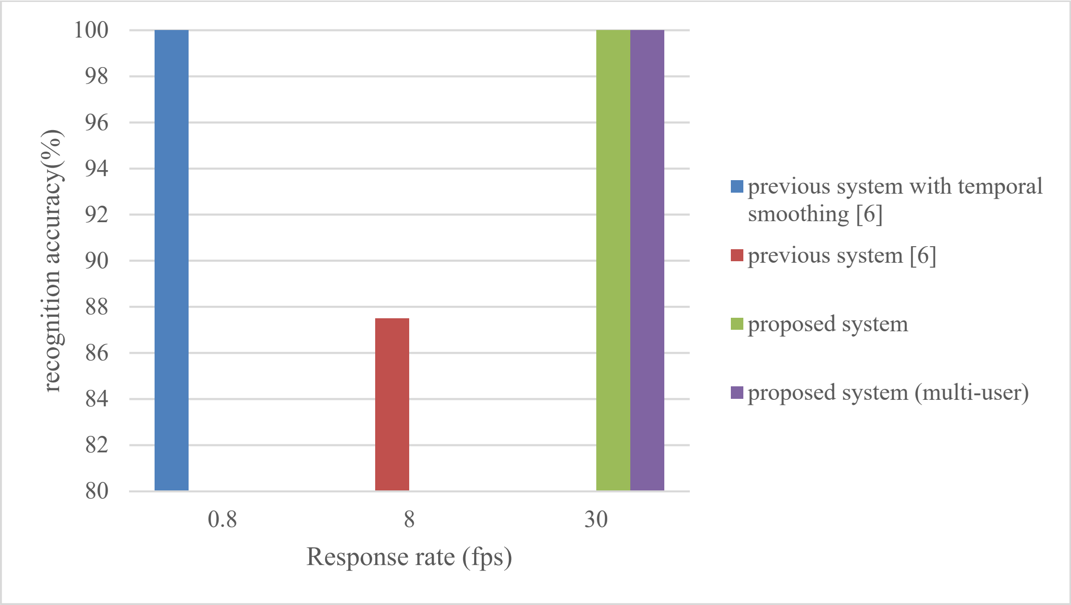


Figure 3.9: Comparison of response rates and recognition accuracies of previous and proposed systems.

3.3.4 User-experience experiment

Both systems (the second system and the first system [6]) are also evaluated by users. A total of 15 participants attended user-experience experiments.

Before starting the experiment, participants were given 3 minutes demonstration on how to use the system with gestures for both systems, followed by one-minute self-directed practice. During this task, the participants were free to ask any questions regarding usage and control of the system.

Following the training, the participant completed the task without interruption. A task consists of the following steps:

S1: Wear 3D glasses and move right hand above the waist for 45cm to start the system.

S2: Rotate models alone in different directions.

S3: Adjust liver's opacity down.

S4: Fusion and selection of vessels

S5: Adjust liver's opacity up.

After finishing the task, participants were moving on to the next system, which again started with a training task. The order of the two systems was randomized.

After completing all trials, the participant responded to a questionnaire. For each of the three basic interactions (rotation, opacity adjustment, fusion, and selection of vessels), participants evaluated four items (intuitive, smoothness, accuracy, and fatigue) in 5 levels. The evaluation decreases as approaching 1 and becomes higher as closer to 5 (not tired is considered as a high score). The definition of four criteria is as follow [18]:

How good was the gesture fitting the visualization result (intuitive)?

How would you evaluate the response time of the system (smoothness)?

How would you evaluate the precise of the system (accuracy)?

How good was the comfort of performing the gestures (fatigue)?

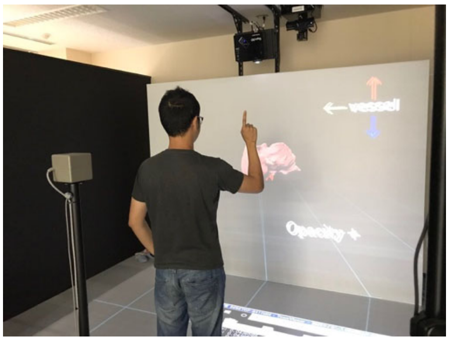
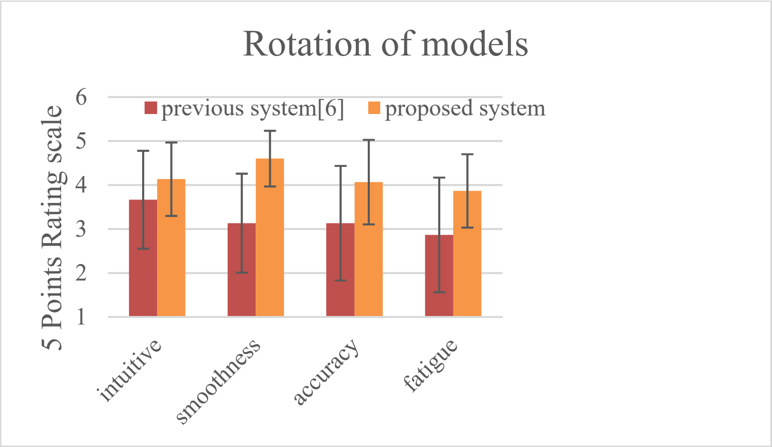
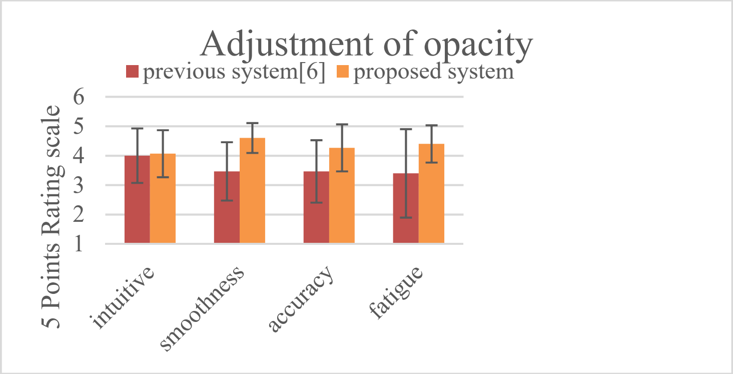


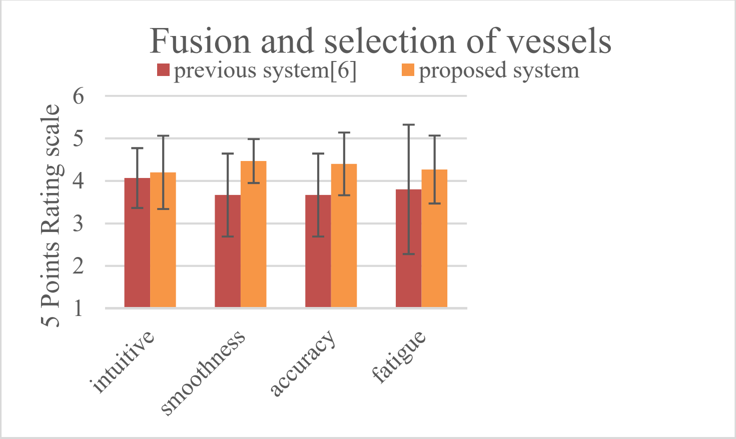
Figure 3.10: A participant adjust liver's opacity up (S5) during the experiment.



(a)



(b)



(c)

Figure 3.11: Mean evaluation scores and the standard errors over all participants between the previous system(red) [6] and the proposed system(orange)for each of the 3 interactions: (a) Rotation of models, (b) adjustment of opacity, (c) Fusion and selection of vessels.

Table 3.3: P-value over all participants for each of the 3 interactions.

Interactions	Subject	P-value
Rotation of models	Intuitive	0.2042
	Smoothness	<0.01
	Accuracy	0.0336
	Fatigue	0.018
Adjustment of opacity	Intuitive	0.8342
	Smoothness	<0.01
	Accuracy	0.0269
	Fatigue	0.024
Fusion and selection of vessels	Intuitive	0.6461
	Smoothness	<0.01
	Accuracy	0.0276
	Fatigue	0.3018

The average evaluation scores are shown in Figure 3.11. It can be seen that the proposed system significantly outperforms the previous system in terms of mean evaluation scores. To confirm there is a statistically significant difference between the proposed system and the previous system, I use the ANOVA (Analysis of Variance) method [14] using a significance level of $\partial = 0.05$. In the test, I have a main null hypothesis as: there is no difference exists between the proposed system and the previous system. The p-value for each evaluation term is shown in table 3. For the first and second interactions (rotation and opacity adjustment), improvements on smoothness ($p < 0.01$), accuracy ($p < 0.05$) and fatigue ($p < 0.05$) are confirmed. For the third interaction (fusion and selection of vessels), improvements on smoothness ($p < 0.01$) and accuracy ($p < 0.05$) are confirmed. Improvements on intuitiveness for all three interactions ($p > 0.05$) and fatigue for the third interaction (fusion and selection of vessels) ($p > 0.05$) could not be confirmed. The reason is that the main contribution of this work is to improve the accuracy and speed of interactions. Further improvements in intuitiveness will be my future work.

3.3.5 Discussion

The field of touchless interaction in surgery has become very active. An excellent survey on existing methods can be found in [11]. In this section, I focus on discussing those approaches most related to my contribution. A practical surgical operation support system wearing HoloLens has been already proposed [13]. HoloLens can detect hand gestures and realize the touchless visualization. The problem for HoloLens-based touchless visualization system is that surgeons must wear the HoloLens during the operation, which is not practical and will limit the surgical performance. Several other touchless systems without wearing glasses have been proposed. A camera-based approach was proposed by Wachs et.al. [10], in which a vision-based technique is used for hand gesture and posture recognition. The algorithm requires a clean background to work, and it is not robust for surgery conditions [11]. The Leap Motion controller represents a revolutionary input device for gesture-based human-computer interaction. It is a stereo camera with 3 infrared LEDs that illuminate the hand over the sensor. But it could not be used as a professional tracking system, due to its rather limited sensory space [12]. Voice commands are also an attractive and potential solution to address this problem. But it is sensitive to other sounds and voices, which are noise for interactions. The advantage of the proposed method is the robustness to the noise. Since my system can obtain depth information, my system can only recognize the action or gesture in a pre-defined specific area (e.g. operation area) and do not response to gestures or actions outside the specific area. The advantage of the proposed method by using Kinect sensor to capture the gesture is its robustness to the noise (other human actions). Jacob's method [16] is like my previous machine learning-based recognition [6], in which I used the histogram of oriented gradients features and a support vector machine (SVM) classifier. Both methods consisted of two processes: feature extraction and classification. Though machine learning-based methods achieved high recognition accuracy, they could not achieve real-time recognition. The main advantage of my proposed method in this paper is that I recognize 3 kinds of hand states and their movements by using the API of the Kinect without processes of feature extraction and classification to realize a real-time interaction. Though the proposed system has limitations: it is not able to realize complicated interactions and lacks the flexibility of interactions, the proposed interactions are enough for touchless visualization control. The originality and novelty of this preliminary study is that I proposed an easy and fast framework to solve this task without doing gesture recognition by ourselves. My system is based on surgeons' advice and requirements. Further experiments involve surgeons will be conducted. It also can be used as education for medical students to help them the visualization of hepatic and vessel structure.

3.4 Multimodal Deep Learning for Accurate Gesture Recognition Using Color and Depth image

3.4.1 MaHG-RGBD: A Multi-angle View hand Gesture RGB-D Dataset

Though the second version (Sec. 3.3) is fast and accurate, it lacks flexibility of interactions because the interaction is only based on the three kinds of hand states and their movements. In order to realize a fast, accurate and flexible touchless visualization system, I developed the third system, whose interaction is based on deep learning-base multi-modal gesture recognition using both color and depth images. In this section, I first present a novel multi-angle view hand gesture RGB-D dataset, recorded with a Kinect. In the first system [6], I had built a dataset, which was collected from 10 participants in advance, each person has 100 pieces of depth images of 9 kinds of hand shapes. It has some limitations, i.e., the number of classes and the amount of data collected. It has only 9 classes and 1000 images per class. And it only contains depth images. The lack of data is one big problem when applying machine learning algorithms for classification. e.g., deep learning algorithms usually required a very large amount of labelled data to obtain acceptable results.

Therefore, I present a new dataset named MaHG-RGBD. The data are collected with RGBD sensors (Kinect) that each cover different views of the hand. A Kinect sensor acquired front-view RGBD videos, and a top-mounted Kinect recorded a pair of RGBD streams. The proposed dataset consists of 25 gestures and one counterpart from a different view. Each class is then recorded by 15 participants and each of them provides 100 images per class by repeating the same hand gesture with slight movements. The main contributions of this dataset are summarized as follows:

- A multi-angle RGB-D dataset with 15 participants performing 25 hand gestures. Not only the front-view (tilted angle=0) but also the tilted view (tilted angle=45 degree) dataset are provided, which can be used when the space is limited.
- Providing a pair of synchronized color and depth well-segmented hand region images. Users can use both or according to their purpose.
- The benchmark on this dataset using deep learning methods. The recognition accuracy for

each gesture and each modality (depth or color) is provided, which will be useful for users to select the robust gestures and image modality (depth or color).

3.4.1.1 The ergonometric design of the dataset recorder system

In the experiments, the distance between the Kinect sensor and the user is 2 meters. The MaHG-RGBD dataset was captured using two RGB-D cameras (Kinect). The Kinect sensor 1 is planted 2.5 meters in height and tilted 45 degrees horizontally. The Kinect sensor 2 is planted 1.2 meters horizontally. Figure 3.12 illustrates the multi-view RGB-D cameras setup. The Kinect FOV (field of view) is 70.6 degrees in the horizontal direction and 60 degrees in the vertical direction [7]. As shown in Figure. 3.12, the experimental setup is in the effective range of Kinect. I capture hand images with a sheet of blue partition as the default background for the color image.

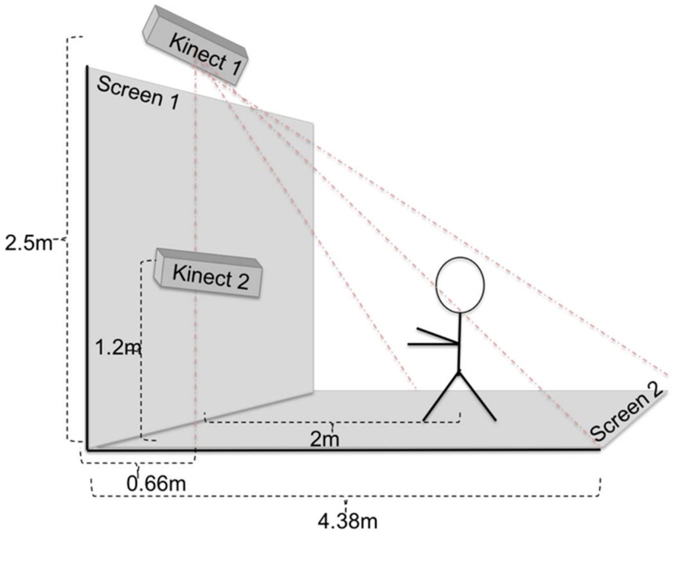


Figure 3.12: Illustration of the dataset acquisition setup.

The dataset includes gestures performed by 15 different subjects. Each subject used their right hand to perform the gestures. During a recording session, each subject provides 100 images per class by repeating the same hand gesture with slight movements.

3.4.1.2 Hand segmentation and pre-processing

I utilize the skeleton tracking provided by the Kinect and the depth information to generate the depth image and color image. First, I acquire a color and depth image of the user, respectively (Figure. 3.13 (a) and 3.13 (b)). Then I do calibration between color and depth camera and using the right hand's joint point as the centre, chip out a 300×300 pixels squire region as an ROI of hand region. The segmented color hand image is shown in the Figure. 3.13 (c). The depth image with a range from d – 30cm and d+5cm is defined as a segmented depth hand region (Figure. 3.13 (d)), where d is the depth of the right hand's joint point. Since the hand image has other regions' pixels with remained as noise, I apply an opening operator and a median filter to remove the noise.

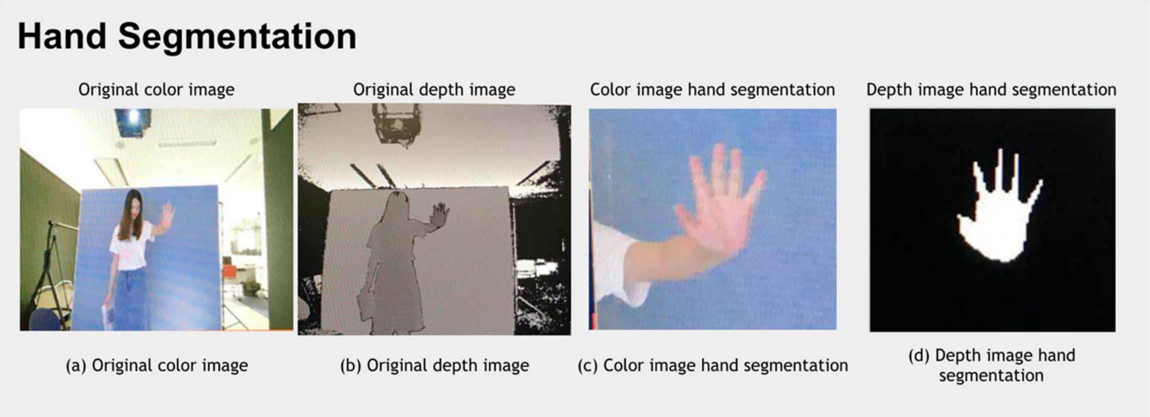

Hand Segmentation
Original color image
Original depth image
Color image hand segmentation
Depth image hand segmentation
(a) Original color image
(b) Original depth image
(c) Color image hand segmentation
(d) Depth image hand segmentation

Figure. 3.13: Sample data for hand image pre-processing

3.4.1.3 Dataset Characteristics

The proposed dataset consists of 25 gestures and one counterpart from a different view. Each class is then recorded by 15 participants, and each of them provides 100 images per class by repeating the same hand gesture with slight movements. The total size of the dataset is 150,000 (2×2×15×25×100) tuples constituted by a depth of the hand region and color of hand region. The size of the hand image is 300×300 pixels. Examples of the multi-angle view hand images of 25 classes in the MaGH-RGBD dataset are shown in Figure 3.14.

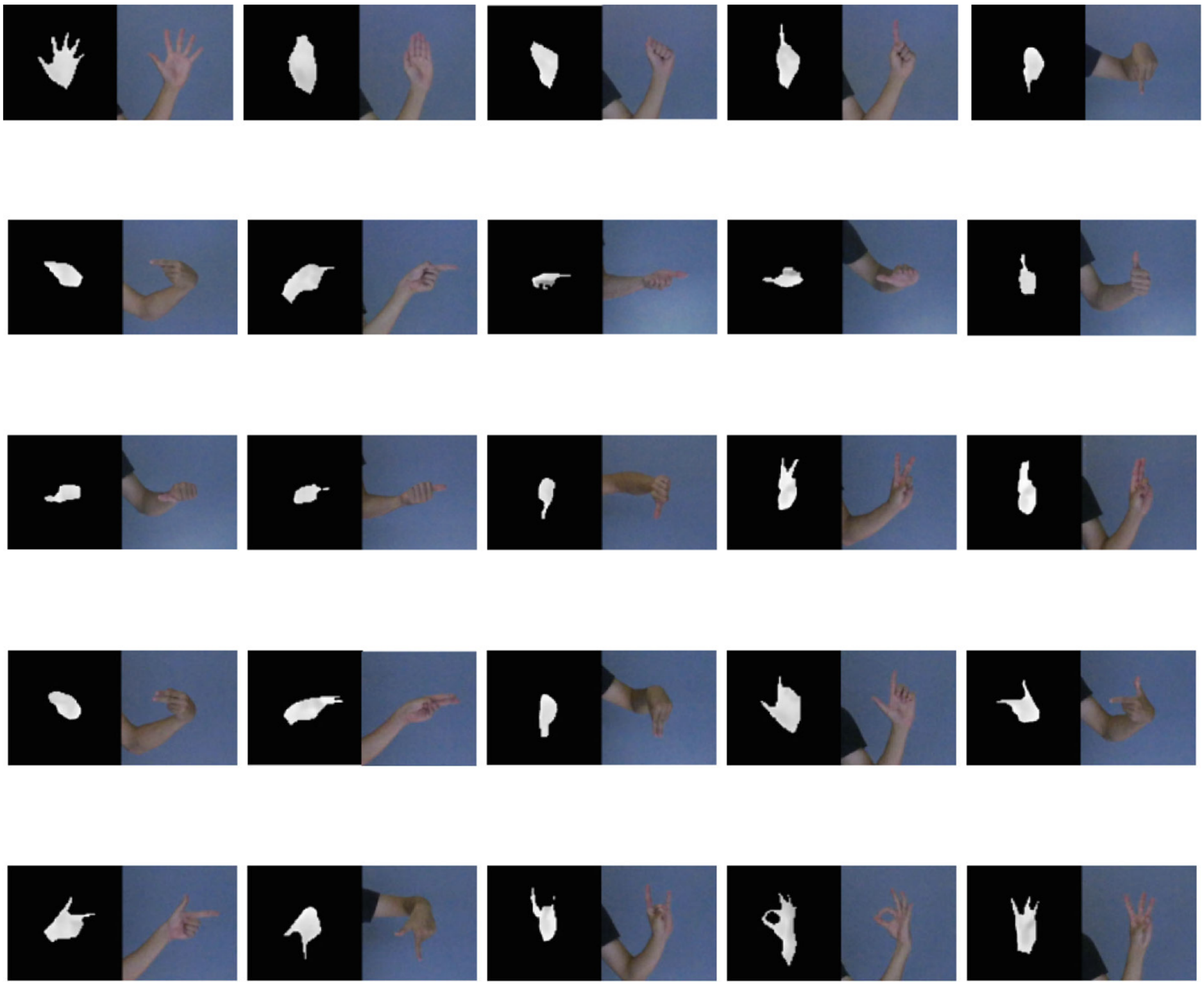


Figure 3.14: Example of 25 classes in the MaHG–RGBD dataset.

3.4.2 Single-modal Deep Learning for Gesture Recognition

Accurate and fast hand gesture recognition is an important requirement for touchless interaction systems. Though the system proposed in Section 3.3 satisfied the requirement, lacks flexibility and it cannot be used for complicated touchless interactions since its freedom is limited. The HOG-based machine learning system has higher freedom and flexibility, but it takes a large computation time, and it cannot work in real-time. In this Chapter, I proposed a new deep-learning-based hand gesture recognition method based on the newly constructed dataset (MaHG-RGBD) and develop a new fast and accurate touchless visualization system for hepatic surgery support. Two well-established networks (LeNet and AlexNet) are used as the baseline networks. I also aim at identifying the architecture that performs best in different image modalities and builds the benchmark of the dataset I proposed in Section 3.4.1.

After that, I have constructed a multimodal deep learning network that adds color information to depth information together and applies the multimodal hand gesture system to touchless medical image visualization.

3.4.2.1 Convolutional neural networks for real-time hand gesture recognition

Since 2012 deep learning-based approaches have consistently shown best-in-class performance in major computer vision tasks [19]. LeNet [21] and AlexNet [19] are selected as the baseline techniques for depth and color hand gestures recognition.

The network architecture of LeNet consists of two convolutional layers, each followed by a pooling layer. And three fully connected layers. The first layer uses 6 kernels and the second 16, both with the same size 5×5. The input of the hand image is 32×32. The output is 25 classes of gestures. The LeNet for hand gesture recognition is shown in Figure 3.15.

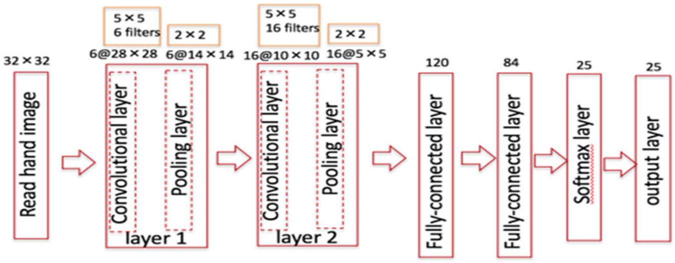


Figure 3.15: LeNet for hand gesture recognition.

The architecture of AlexNet is summarized in Figure 3.16. It contains eight learned layers – five convolutional and three fully connected layers. The input of the hand image is 224×224. The output is 25 classes of gestures.

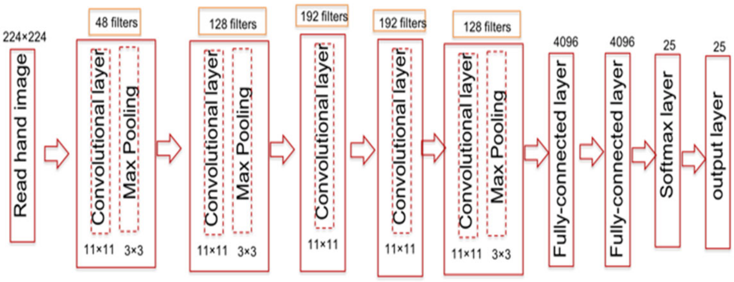


Figure 3.16: AlexNet for hand gesture recognition

3.4.2.2 Recognition benchmark results

As a verification method, I used 15-fold cross-validation. 14 persons were used as training data, 1 person was used as test data. The number of training samples is 35,000 and the number of test samples is 2,500 for all 25 classes, respectively. I repeated it 15 times in total and verified the results of all cases. It can be seen in Table 3.4 that the average accuracy of classification for LeNet and AlexNet.

Table 3.4: Benchmark on MaGH RGB-D Dataset

Classifier	Data types	Test Accuracy
LeNet [21]	Tilted Angle = 0 Depth image	89.98±3.67
	Tilted Angle = 45 Depth image	84.56±7.09
	Tilted Angle = 0 Color image	79.06±6.53
	Tilted Angle = 45 Color image	72.66±13.51
AlexNet [19]	Tilted Angle = 0 Depth image	95.41±2.79
	Tilted Angle = 45 Depth image	92.32±5.05
	Tilted Angle = 0 Color image	88.42±6.23
	Tilted Angle = 45 Color image	79.58±11.97

Table I illustrate the accuracy results from depth image titled angle = 0, depth image titled angle = 45, the color image titled angle = 0, and color image titled angel = 45 using LeNet and AlexNet. From the results, the AlexNet significantly improves the results for all tested datasets. For example, the overall test accuracy is increased from 89.9% to 95.4%, 84.5% to 92.3%, 79.06% to 88.4%, 72.6% to 79.58% for the depth titled angle = 0, angle = 45, and color titled angle = 0, angle = 45, respectively. We can also find that the recognition results of depth images are higher than the results of color images in both methods and different angles of view. This suggests that the depth image is more appropriate for the hand classification task.

To examine the results in more detail, the benchmark comparative results for 15 participants are shown in Figure 3.17 and Figure 3.18 for LeNet and AlexNet, respectively.

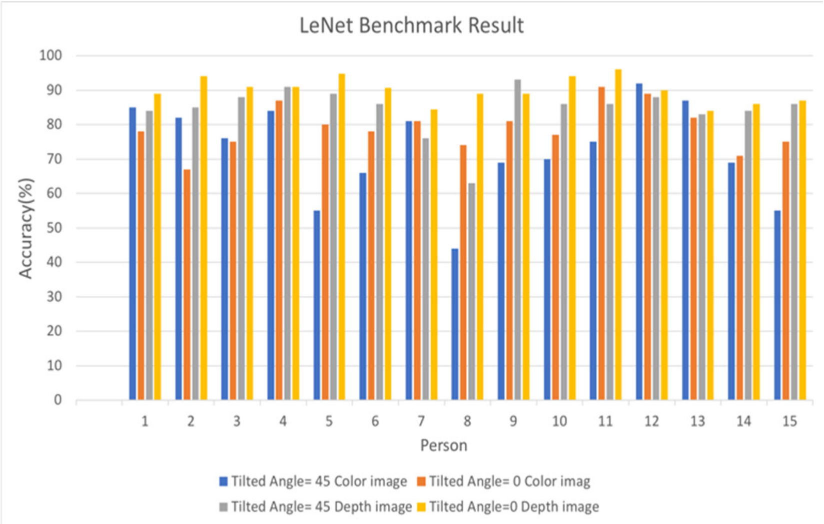


Figure 3.17: LeNet benchmark result for 15 participants.

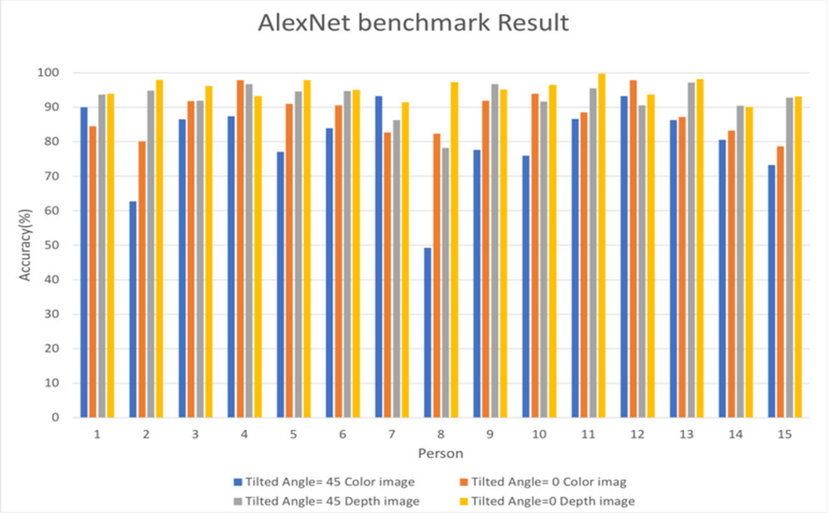


Figure 3.18: AlexNet benchmark result for 15 participants.

3.4.3 Multimodal Deep Learning for Accurate Gesture Recognition Using Color and Depth image

3.4.3.1 Multi-modal deep learning network

The proposed method employs multimodal deep learning gesture recognition consisting of two types of images, i.e., depth and color images. This method aims to improve gesture recognition accuracy using depth and color images. Figure 3.19 shows the network architecture of multimodal deep learning. The depth and color images are input into the two AlexNet learned with the distance and color images, respectively. The average value of the network output is taken as the final output. The output is computed with the following Equation (3.2):

$$y = \arg\underset{k}{max}\left(\frac{yd_k + yc_k}{2}\right) \tag{3.2}$$

where yd_k and yc_k are the probability that the output of the depth and color image network, respectively, belongs to class k.

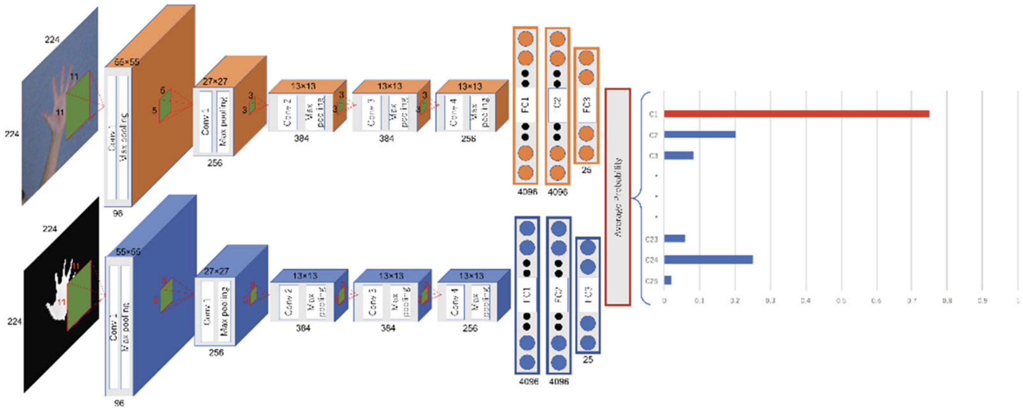


Figure 3.19: Multimodal deep learning network for hand gesture recognition.

3.4.3.2 Experiment results

I use a 15 -fold cross-validation method. A total of 14 persons were used as training data, and one person for testing. Twenty percent of the training data are selected randomly as the validation set. I repeat the process 15 times and verify the results of all the cases. The mean accuracy of 15 times is used as a measure of recognition accuracy, which is defining in Equation (3.3), where Acc_{k} is the accuracy of k-th experiment.

$$\text{accuracy} = \frac{\sum_{k=1}^{15} (\mathrm{Acc}_{k})}{15} \tag{3.3}$$

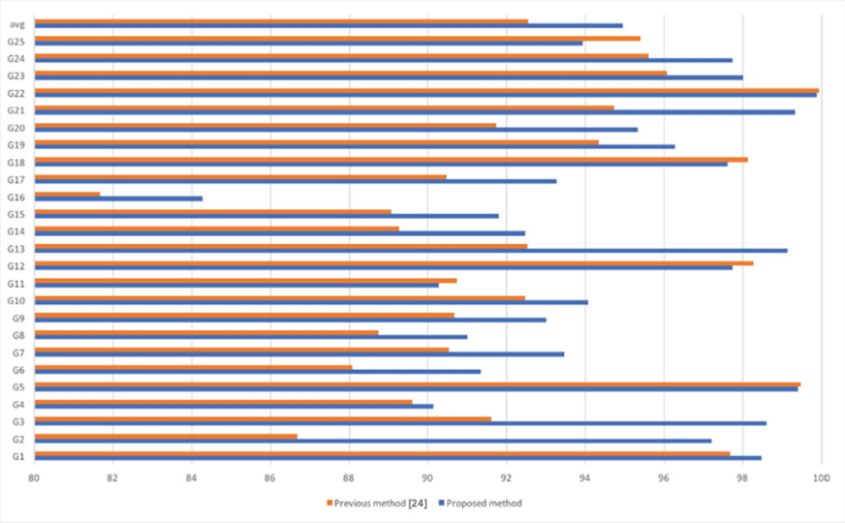


Figure 3.20: Comparison of the recognition result of the proposed method and previous method [24]. The orange graph is the previous method, and the blue graph is the proposed method.

Figure 3.20 presents a comparison between the results of recognition accuracy obtained from the previous research [24] and those obtained using the proposed method. In the graph, results obtained using the previous and a proposed method are denoted in orange and blue, respectively. Most gesture recognition accuracy has been improved. In the recognition using only the depth image, the accuracy is 92.53%, whereas, in the recognition using the depth and color images, the accuracy is 94.94%, showing a 2.41% improvement in the accuracy. Using a color image, the accuracy is significantly improved for gesture recognition accuracy. However, poor accuracy is remarked for only the depth image. Since the recognition time consumption is 0.0146 s, the implementation in real-time is similar.

Table 3.5 lists the average confusion matrix of the proposed method. Misrecognition of a specific gesture is remarked when gesture recognition accuracy is less than 90%. Thus, the gesture recognition is more robust, if the gesture in the touchless interface does not use an easily misrecognized gesture.

Table 3.5. The average confusion matrix of the proposed method. Lager misclassification errors (> 5) are indicated in red and zero values and not included.

Truth	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18	G19	G20	G21	G22	G23	G24	G25
G1	98.47	0.07									0.20	0.07				0.60			0.27		0.13			0.13	0.07
G2		97.20	0.07	0.07	0.13	0.40			0.07		0.13	0.07				1.20		0.20	0.27			0.13		0.07	
G3			98.60	0.07				0.20								0.40		0.13						0.60	
G4			1.13	90.13						2.80					5.13	0.73							0.07		
G5			0.13		99.40				0.07									0.40							
G6	0.07					91.33	0.13		0.07		0.53					6.87	0.07		0.87	0.07					
G7						0.07	93.47					5.87				0.07	0.53								
G8			5.13					91.00	2.07			0.27													1.53
G9		0.07	4.47					0.13	93.00		1.80							0.53							
G10						0.27				94.07	0.07		1.07	0.67	3.13			0.20		0.40					0.13
G11			2.27	0.07		0.13				1.33	90.27		0.07			5.87									
G12	0.13	0.33					0.13	0.33				97.73	1.27					0.07							
G13								0.07	0.47				99.13					0.33							
G14	1.67			0.13						0.53				92.47	0.07								4.73	0.07	0.33
G15				2.80					0.07	3.07			0.07	0.80	91.80								1.33		0.07
G16						8.13	0.07				0.67		5.80		0.07	84.27				0.67		0.33			
G17	0.07						1.13	0.20			0.07	4.47					93.27				0.80				
G18					0.20								2.20					97.60							
G19						0.40			0.07		0.20					1.93			96.27	1.13					
G20						0.80					0.07								3.73	95.33					0.07
G21							0.60	0.07													99.33				
G22											0.07							0.07				99.87			
G23				0.20										0.80					0.40				98.00		0.60
G24	0.80	0.27	1.13																				0.07	97.73	
G25				0.27						2.60			1.87	0.67									0.67		93.93

3.5 Chapter Summary

In this Chapter, I proposed a real-time gesture recognition system for a touchless hepatic surgery support system. I have proposed four versions.

In the first version, I used HOG as features and SVM as a classifier to recognize 9 kinds of hand gestures from the depth images, the average recognition accuracy is found to be 87.5% with a speed of 8fps. Though the HOG-based machine learning method can recognize various hand gestures with reasonable accuracy, they could not achieve real-time recognition.

I describe the second version in Section 3.3. the system uses a Kinect sensor to acquire three kinds of hand states and track hand their movements. Based on these states and their movements, I designed a range of hand gestures, and finally, four kinds of operations are available using touchless gestures to visualize 3D hepatic anatomic models in real-time. Although this version is a prototype, the preliminary result is encouraged. The experiments demonstrated that the recognition rate of 100% has been achieved in the proposed system even at the frame rate of 30fps (real-time). The use-experiment showed that the proposed system significantly improved the smoothness, accuracy, and fatigue (except the interaction of vessel selection and fusion). This proposed system also can be used as an education system for medical students to help them understanding the anatomical structures of humans. I also realized an interaction robust to noise (second version). In addition to hand state and movements, I also used depth information (predefined range) to constrain the users. The actions or gestures out of the predefined range (2.5m ~ 3.5m) were considered as noise and the system only responded to the gestures in the predefined range (e.g. operation range). So that the system can only respond to the gestures of the surgeon, which is important especially in the surgery room.

Further improvements on intuitiveness are in the third version of the system (Section 3.4). Though the second version of only 4 operations, I develop a deep learning technique for recognition of various hand gestures to increase the degree of freedom of operations and achieve more flexible touchless visualization. Since deep learning usually required a very large amount of labelled data to obtain acceptable results, I built a new multi-view RGB-D dataset (MaHG-RGBD) with 15 participants performing 25 hand gestures [24]. Not only the front-view, but also the tilted view (titled angle = 45 degrees) dataset are provided, which can be used when space is limited especially in the surgery room. After building the dataset, I use AlexNet to recognize hand gestures and select 9 robust hand gestures for touchless visualization of 3D medical images. Based on the new dataset I primarily focus on selecting robust hand gestures for the touchless visualization system. A rapidly responding and flexible Kinect-based touchless visualization has been realized. I also propose a multimodal deep learning method to perform recognition using color and depth images. The multimodal system achieves better real-time robust recognition than conventional methods.

Bibliography

1. Kaibori M, Chen YW, Matsui K, Ishizaki M, Tsuda T, Nakatake R, Sakaguchi T, Matsushima H, Miyawaki K, Shindo T, Tateyama T, Kwon AH: Novel Liver Visualization and Surgical Simulation System. J Gastrointest Surg. 17, pp.1422-1428, 2013.
2. Tateyama T, Kaibori M, Chen YW et al.: Patient-specified 3D-visualization for Liver and Vascular Structures and Interactive Surgical Planning System. Medical Imaging Technology, 31, pp.176-188, 2013 (in Japanese).
3. Gallo.L.: Controller-free exploration of medical image data: experiencing the Kinect, National Research Council of Italy Institute for High Performance Computing and Networking, 2011.
4. Yoshimitsu K, Muragaki Y, Iseki H. et al.: Development and Initial Clinical Testing of “OPECT”: An Innovative Device for Fully Intangible Control of the Intraoperative Image-Displaying Monitor by the Surgeon. Neurosurgery. Suppl 1, pp.46-50, 2014.
5. Ruppert GC, Coares C, Lopes V, et al. Touchless gesture user interface for interactive image visualization in urological surgery. World J. Urol. 30, pp.687-691, 2012.
6. Jia-Qing Liu, Ryoma Fujii, Tomoko Tateyama, Yutaro Iwamoto, Yen-Wei Chen, Kinect-Based Gesture Recognition for Touchless Visualization of Medical Images, International Journal on Computer Electrical Engineering, Vol.9, pp.421-429, 2017.
7. Microsoft, Kinect for Windows Software Development Kit (SDK) 2.0 URL: https://www.microsoft.com/en-us/download/details.aspx?id=44561
8. Shotton et al., ‘‘Real-Time Human Pose Recognition in Parts from a Single Depth Image,‘‘ Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), IEEE CS Press, 2011, pp. 1297-1304.
9. Sugimoto M et al., Augmented Tangibility Surgical Navigation Using Spatial Interactive 3-D Hologram zSpace with OsiriX and Bio-Texture 3-D Organ Modeling. 2015 International Conference on Computer Application Technologies, pp.189-194, 2015.
10. Wachs JP, Stern HI, et al., A gesture-based tool for sterile browsing of radiology images. Journal of the American Medical Information Association 15(3):321-323.
11. Mewes, A., Hensen, B., Wacker, F. et al. Touchless Interaction with software in Interventional Radiology and Surgery: A Systematic Literature review. Int J CARS (2017)

12. Guna, J., Jakus, G., Pogačnik, M., Tomažič, S., & Sodnik, J. (2014). An Analysis of the Precision and Reliability of the Leap Motion Sensor and Its Suitability for Static and Dynamic Tracking. *Sensors (Basel, Switzerland)*, *14*(2), 3702–3720.
13. Sugimoto M, HoloEyes sharing Mixed Reality for surgical navigation URL:
14. https://wired.jp/waia/2017/14_maki-sugimoto/
15. Statistical Solutions (2013). Retrieved from http://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/anova/
16. Jacob MG, Wachs JP, Context-based hand gesture recognition for the operating room, Pattern Recognition Letters, Vol. 36, p 196-203, 2014.
17. Magnetic 3D URL: http://www.magnetic3d.com/
18. Soutschek S, Penne J, Hornegger J, Kornhuber J (2008) 3-d gesture-based scene navigation in medical imaging applications using time-of-flight cameras. In: IEEE computer society conference on computer vision and pattern recognition workshops, 2008. CVPRW'08.
19. Krizhevsky, A. Sutsjever, I. and Hinton, G. "ImageNet classification with deep convolutional neural networks", Adv. Neural Inf. Process.Sys., pp9, 2012.
20. Shukla, D. Erkent, O. and Piater, J. "The IMGH dataset: A Multi-View Hand Gesture RGB-D Dataset for Human-Robot Interaction"
21. Lecun, Y. Bottou, L. Bengio Y. and Haffner, P. "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324.
22. Liu, J.Q., Tateyama, T., Iwamoto, Y., Chen, YW.: A Preliminary Study of Kinect-Based Real-Time Hand Gesture Interaction System for Touchless Visualization of Hepatic Structure in Surgery. Medical Imaging and Information Sciences, Vol. 36, no. 3, pp. 128-135, 2019.
23. Liu, J.Q., Tateyama, T., Iwamoto, Y., Chen, YW.: Kinect-Based Real-Time Gesture Recognition Using Deep Convolutional Neural Networks for Touchless Visualization of Hepatic Anatomical Models in Surgery. KES-IIMSS-18 2018. Smart Innovation, Systems and Technologies, vol 98. Springer, Cham.
24. Liu, J.Q., Furusawa K., Tsujinaga S., Tateyama, T., Iwamoto, Y., Chen, Y.-W., 2018.: MaHG-RGBD: A Multi-angle View Hand Gesture RGB-D Dataset for Deep Learning Based Gesture Recognition and Baseline Evaluations. Proc. of IEEE ICCE2019, Las Vegas, USA, Jan. 11-13, 2019.

Chapter 4

Multimodal Image Generation for Improving Single-modal Posture Recognition

4.1 Introduction

Human posture classification has recently received much attention for its wide potential applicability in areas such as: augmented reality, human-computer interaction, and rehabilitation [1]. Based on their input type (RGB images or depth images), human pose estimation tasks can be divided into two classes. The biggest difference between them is that pixels in the RGB image record the color information of the subject, while pixels in the depth image record the distance between the subject and the cameras. Pose and gesture estimation from an RGB image has been realized by various approaches. For instance, Priya [2] proposed a CNN for classifying multi-view human pose datasets. Pinto [3] proposed a CNN-based static hand gesture recognition method. The other source of input is depth information. Nishi [4] proposed an efficient generation of human depth images with body part labels and verified the constructed dataset using a fully convolutional network (FCN). Wang [5] proposed a human pose recognition based on the fusion of local difference of depth feature (LDoD) and directional gradient of depth feature (DGoD) features. Comparing with RGB images, depth images provide distance information that is important to overcome the confusion of body parts and occlusions.

Generative adversarial networks (GANs) have been explored in various posture estimation tasks. Wan et al. [6] proposed the VAE and GAN networks (Grossing Nets) to estimate 3D hand gestures from single depth images. He et al. [7] developed a framework that combines GANs and style transfer for depth hand image synthesizing from 3D hand gestures. Unlike [8], which tends to build the correlation between depth image and 3D hand joints. This paper generates the estimated depth image from the color image through hybrid loss GANs.

Although depth images show a great advantage, depth cameras are costlier and less widely used than color cameras. To avert the difficulty of acquiring depth images, I generate estimated depth images for improved human posture recognition based on color images, which is inspired by my previous work [8]. This is a further key distinction with the existing posture recognition methods, which enable the sub-sequence stage to produce accurate classification results.

In this Chapter, I focus on the image-based classification of posture recognition. I propose a novel deep learning method using a two-stage CNN architecture. The first stage is to estimate depth posture using the generative adversarial networks (GANs). In the second stage, I build two-stream CNNs to learn feature representation for the input RGB and their corresponding estimated depth image from the first stage, which are then combined with feature fusion. This architecture is similar to other recent multi-stream CNNs [9]. However, in this work, I consider the relationship between the RGB image and the depth image, to avert the difficulty of acquiring depth images. The main advantage of this two-stage architecture is that it exploits the RGB image and the depth estimation at training time and recognizes the posture directly from only the RGB image at the testing time, with an average processing rate of 34.44ms per frame on the novel dataset.

The contributions of this Chapter are as follows: (1) a generated adversarial network (GAN)-based feature augmentation subnetwork for estimated depth posture generation, which improves the performance of posture recognition; (2) a hybrid loss function for the generation module, which generates estimated depth posture image while capturing the high-level features and recovering the sharp depth discontinuities; (3) a novel dataset of 13800 pairs of color and depth human pose images, which is used for depth map estimation from the single color image. The dataset is available at: http://media.ritsumei.ac.jp/iipl/database/pose/.

The remainder of this Chapter is organized as follows. The details of the proposed network are described in subsection 4.2. Subsection 4.3 presents the experimental results. The confusions are present in Subsection 4.4

4.2 Proposed Method

This study proposes a high-accuracy human posture recognition system using RGB color images alone. The structure of the proposed method is shown in Figure 4.1. My approach has two main stages: generation of an estimated depth image from a color input image, and recognition of the human posture using both input color image and its estimated depth image. The first stage is realized by an improved Pix2pix network, and second stage is realized by a two-stream CNN network. The first stage also includes a hybrid loss function that generates estimated depth posture images while capturing the high-level features and recovering the sharp depth discontinuities.

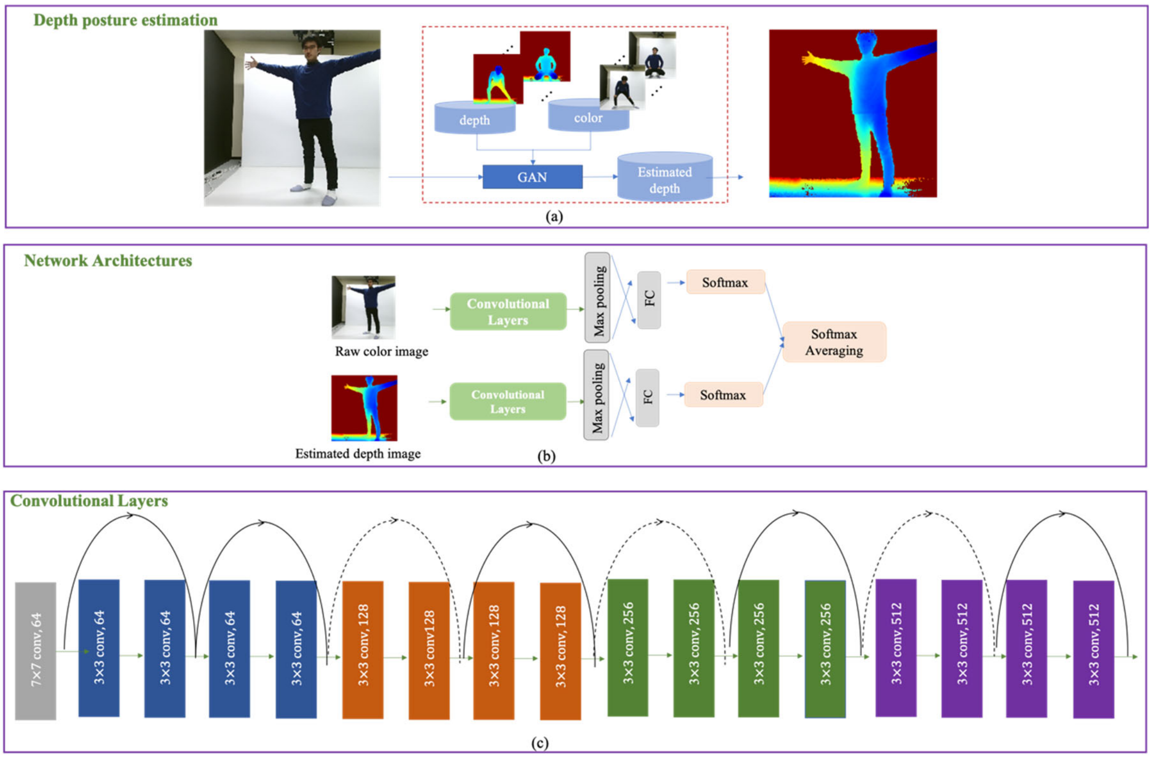


Figure 4.1: Overview of the proposed method. (a) An example of Depth posture estimation using the generative adversarial network. (b) Different approach for fusing information from raw color image and estimated depth image. Each green box represents a convolutional stream. The left part is late fusion, and the right part is committee fusion. (c) Details of the convolutional stream.

4.2.1 The generation networks

The generative network is shown in Figure 4.2. The network architecture is based on Image-to-Image Translation with Conditional Adversarial networks (Pix2pix) [10-12]. GANs are composed of a generator G and a discriminator D. The goal of GANs on depth estimation task is to learn a mapping from the input RGB image x_i to the target image y_i. During training, G aims to deceive the D by approximating the real data y_i distribution to generate the image G(x_i), whereas D tries to distinguish between real image y_i and fake image G(x_i). The detailed parameters of the improved single channel generator network and discriminator network are shown in Figure 4.3.

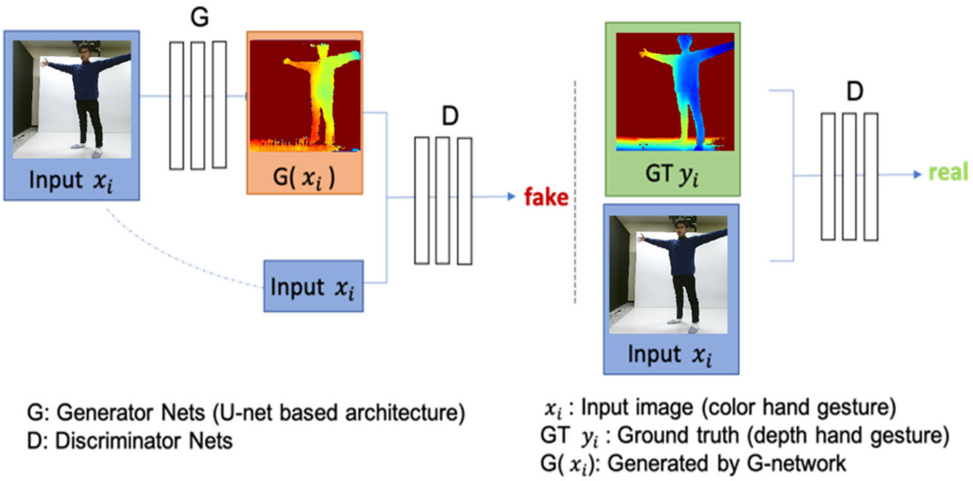
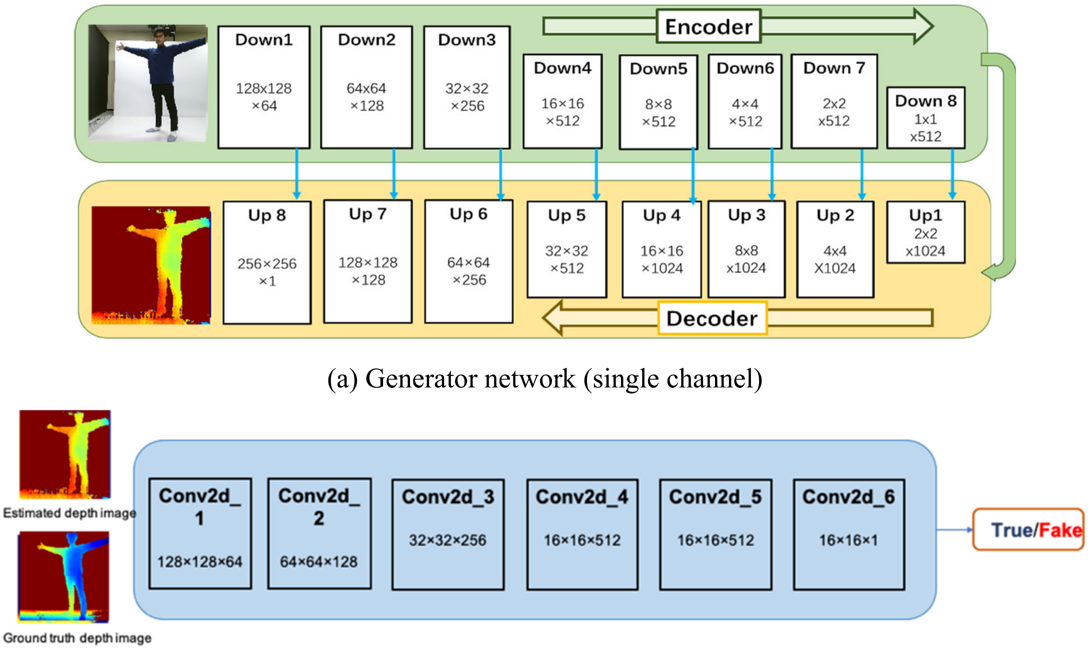

G
Input x_i
G(x_i)
D
fake
Input x_i
D
GT y_i
Input x_i
real
G: Generator Nets (U-net based architecture)
D: Discriminator Nets
x_i : Input image (color hand gesture)
GT y_i : Ground truth (depth hand gesture)
G(x_i): Generated by G-network

Figure 4.2: Structure of my generative adversarial network



(a) Generator network (single channel)

(b) Discriminator network.

Figure 4.3: Parameters of the improved generator network (single channel) and discriminator network.

For generator G and its discriminator D, the adversarial loss can be written as

$$L_{GAN}(G,D) = E_{x\sim p_{\text{data}}(x)}[\log D(x \mid y)] + E_{G(x)\sim p_x G(x)}[\log(1 - D(G(x) \mid y))] \qquad (4.1)$$

I employ L1 loss on the generator to enforce the pixel-wise consistency between generated and real image. The L1 loss is defined as

$$L_{L1} = E_{x,y\sim p_{\text{data}}(x,y)}[\| y - G(x) \|]] \qquad (4.2)$$

The overall objective in the Image-to-Image Translation with conditional adversarial networks can be expressed as

$$G^* = \arg \min_G \max_D L_{GAN}(G,D) + \lambda L_{L1}(G) \qquad (4.3)$$

Let $\{x_i\}_{i=1}^N (x_i \in X)$ and $\{y_i\}_{i=1}^N (y_i \in Y)$ be the color and real depth pose images respectively. My goal is to learn a mapping function between two domains x and y-based training a dataset (pairs of x and y). The generator aims to minimize the loss value, and the discriminator aim to maximize the loss value. (Setting λ= 100).

For the loss value of the generator network, we add perceptual loss [13],[15], gradient loss [14] to the loss of the generator's final output.

The perceptual loss is estimated by using the L2 norm between the feature maps from the predicted and ground truth depth maps.

$$\ell_{\text{perceptual}} = \frac{1}{CHW} \| f(G(X)) - f(y) \|^2 \qquad (4.4)$$

$f(G(X))$ and $f(y)$ are the activations of the 7th convolutional layer of the generator network of the shape C×H×W based on the predicted depth map and ground truth depth map, respectively.

To recover the sharp depth discontinuities and smooth gradient changes in the predicted depth images, I also consider

$$\ell_{\text{gradient}} = \frac{1}{n}(|\nabla \text{x}(\text{G}(x) - y)| + |\nabla \text{y}(\text{G}(x) - y)|) \qquad (4.5)$$

where, ∇x and ∇y represent the depth derivatives in x and y directions, respectively. The ℓ_{gradient} gives the L1 difference between the predicted logarithmic depth derivatives in x and y directions and the ground truth logarithmic of their depth derivatives.

The total loss can be calculated by solving the following Equation (4.6):

$$\ell_{GANs} = \arg \min_G \max_D L_{GAN}(G,D) + \lambda L_{L1}(G) + \lambda_1 \ell_{\text{perceptual}} + \lambda_2 \ell_{\text{gradient}} \qquad (4.6)$$

Where, λ_1 and λ_2 are the wight of perceptual loss and gradient loss. The effect of varying these 3 losses on the final output will be evaluated in a study of the estimated depth posture.

4.2.2 The classification networks

My classifier network is shown in Figure 4.1(b). this stage consists of two CNNs with the same architecture. One stream extracts features from the raw color image; the other stream extracts features from the estimated depth image. For a baseline, I build an 18-layer ResNet [16], which has achieved great performance on image classification. The Figure 4.1 (c) presents the detail of my ResNet. I adopted batch normalization after each convolution and before activation (ReLU). Figure 4.1(b) presents the ways for fusing information from RGB and estimated depth images. I employ the independent convolutional layers for RGB and estimated images. The outputs of the max pooling layers are concatenated and fed into a two-way shared fully connected layer with softmax to compute a cross-entropy classification loss.

4.3 Experiment and Evaluation

This section evaluates the performance of the proposed model on two benchmark datasets. One is the released novel human pose dataset. The second is the public OUHANDS hand gesture dataset. In particular, I discuss the implementation details, list the standard metrics for comparing the generative models, compare the estimated depth generations with different losses, and summarize the classification results.

4.3.1 Experiment on human pose dataset

4.3.1.1 The ergonomics design to the dataset recorder system

The first dataset is the novel human poses dataset, which contains 13,800 samples of paired color-and-depth of 6 subjects with 15 postures obtained by Kinect V2 sensor [19]. In the experiments, the Kinect V2 sensor was displaced horizontally by 1.2 meters. Figure 4.4. shows the dataset acquisition setup. In this study, the distance information in the range from 1.1 m to

2.375 m is converted into grayscale. The depth images obtained by the Kinect V2 sensor are considered as the ground-truth depth map. In order to make an easy visual understanding, I use a pseudo color to represent the depth images in Figure 4.4 I set the minimum value of the human body-field to 0 and the maximum value to 1 and normalized each picture. The pixel values larger than the maximum value are set to 1, and the pixel values smaller than the minimum value are set to 0.

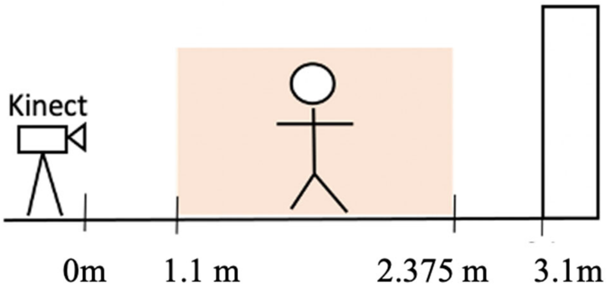
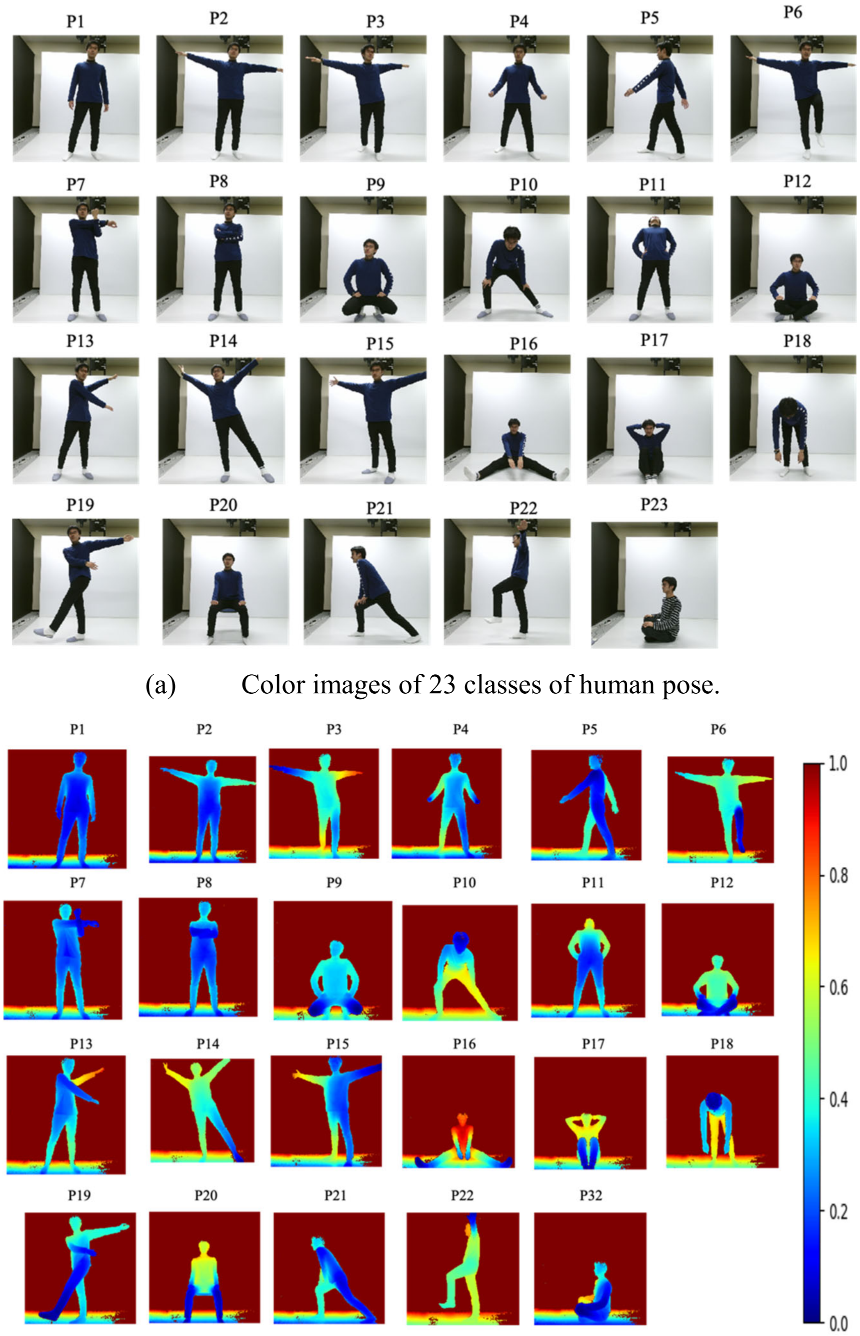


Figure 4.4: Illustration of the dataset acquisition setup.

4.3.1.2 Human poses dataset

This dataset contains 23 poses acquired from six subjects. The resolution of the image taken with Kinect V2 is 1920 × 1080 for color image and 512 × 424 for depth image. The total size of the dataset is 13,800 (23×6×100) tuples of depth and color images. The dataset is split into training, validation, and testing sets with 2300, 2300, and 9200 images, respectively. I use this dataset to evaluate the proposed method for both depth estimation and pose recognition. The mentioned 23 poses are demonstrated in Figure 4.5.



(a) Color images of 23 classes of human pose.

(b) Depth images of 23 classes of human pose (with color bar)

Figure 4.5: Paired color (a) and depth (b) images of the 23 human-pose classes compiled in the dataset.

4.3.1.3 Implementation details

The generation network was trained for 100 epochs with a batch size of 16. When I apply perceptual loss, as the weights in the network are not pre-trained, I first trained the network using the MSE loss for 10 epochs. Next, the ground-truth and estimated depth maps were passed through the network, and the loss was measured using the feature maps in the penultimate layer of the encoder. During the training, 20 pieces were randomly selected from each pose of the training set (using 1840 images in total). In addition, the input image was clipped to a resolution of 256 × 256.

The classification network training was performed for 200 epochs by the SGD optimizer with a batch size of 64 and an initial learning rate of 0.001. All results were generated on an NVIDIA GeForce RTX 3070 GPU and the network was implemented using the PyTorch library.

4.3.1.4 Evaluation of the estimation depth map

I qualitatively compare the performances of the baseline Pix2pix [11] and my approach with different combinations of loss functions. The evaluation set was excluded from training.

Denoting the total number of valid pixels (non-background pixels) in each evaluation by P, I assessed the performances by the following accuracy measures, which are commonly used in related studies [17]:

1) Mean relative error: $\mathrm{REL} = \frac{1}{p}\sum_{i=1}^{p} \frac{\|d_i - g_i\|_1}{g_i}$
2) Thresholder accuracy: Percentage of d_i satisfying $max\left(\frac{d_i}{g_i}, \frac{g_i}{d_i}\right) = \delta <$ threshold

Using these popular measures, I can compare the depth accuracy of different methods from multiple perspectives. The threshold in the second measure was varied as $thr_1 = 1.25$, $thr_2 = 1.25^2$, $thr_3 = 1.25^3$. The assessment metrics are quantitative compared in Table 4.1 and are visually compared in figure 4.6. Pix2pix (setting λ_1 = 0, λ_2 = 0 in Eqn.4.4-4.6), Perceptual (setting λ_1=0.5, λ_2 = 0 in Eqn. 4.4-4.6, Gradient (setting λ_1= 0, λ_2=0.5 in Eqn. 4.4-4.6) and Hybrid loss (setting λ_1 = 0.25, λ_2 = 0.25 in Eqn. 4.4-4.6 methods are estimated. The proposed hybrid loss significantly produced better results and more precise details than the original Pix2pix method.

Table 4.1: Quantitative comparison of depth map estimation.

Method	REL	Thr1	Thr2	Thr3
Pix2pix [11]	0.241	0.843	0.888	0.923
Perceptual [15]	0.304	0.844	0.872	0.904
Gradient [14]	0.297	0.861	0.895	0.923
Hybrid loss	**0.137**	**0.864**	**0.920**	**0.947**

The visual results on the pose dataset for the different approaches are shown in Figure 4.6. I have the following common findings. (1) Pix2pix, gradient loss, and perceptual loss have grid-like artifacts at the pixel level, which leads to an unsatisfactory visual quality. (2) Compared to other methods, my hybrid loss produces visually realistic images with more accurate detail. I believe that the hybrid loss helps the network to predict more accurate images by incorporating additional constraints to enforce appearance consistency between predicted and ground-truth images.

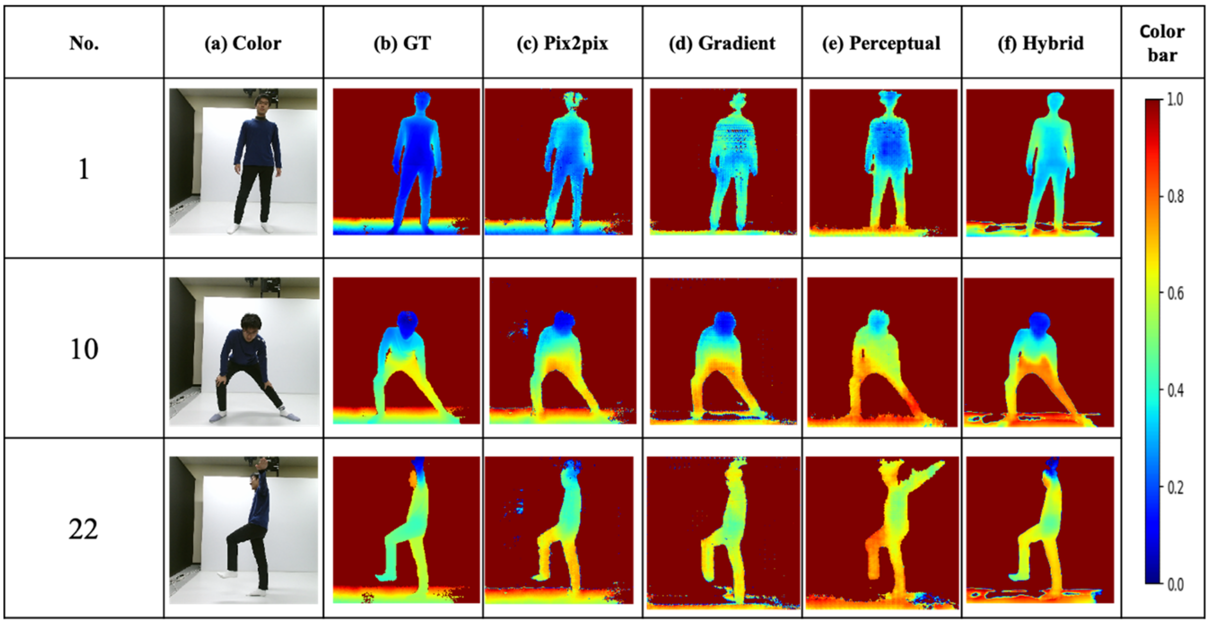


Figure 4.6: Visual comparisons on pose dataset: (a) Color images (input), (b) ground truth, (c) pix2pix results, (d) gradient loss results, (e) perceptual loss results, (f) Hybrid loss results.

4.3.1.5 Evaluation of the pose recognition

This subsection demonstrates the effectiveness of the proposed method on human posture recognition. I first perform ablation experiments to show the effect of each key component. The experiments include the baseline using RGB image only, methods using estimated depth image with different loss functions, and two-steam architecture with both color and estimated depth images. The results are summarized in Table 4.2. As shown in Table 4.2, compared to the baseline using RGB image only, the methods using estimated depth image achieved better results. Compared with existing loss functions, the proposed hybrid loss improved the recognition accuracy to 0.898 from 0.863 (pix2pix loss [11]), 0.816 (gradient loss [14], and 0.870 (perceptual loss [15]), respectively. The accuracy was further improved to 0.967 using a two-stream architecture with both color and estimated depth images (the proposed method).

Table 4.2: Ablation experiment for proposed method.

Color	Estimated Depth	Loss	Acc
V			0.778
	V	pix2pix loss	0.863
	V	perceptual loss	0.870
	V	gradient loss	0.816
	V	hybrid loss	0.898
V	V	hybrid loss	**0.967**

Table 4.3: Comparison of the proposed method and the-state-of-the-art on the pose test set.

Model	Input	Acc	Precision	Recall	F1 Score	Time (ms)
Pinto. [3]	RGB	0.778	0.778	0.778	0.750	15.00
Liu. [8]	RGB	0.863	0.874	0.863	0.850	30.95
Zheng [14]	RGB	0.816	0.855	0.816	0.793	30.95
Kumari [15]	RGB	0.870	0.902	0.870	0.862	30.95
Thomas [22]	skeleton	0.812	0.812	0.786	0.790	14.73
Proposed method	RGB	**0.967**	**0.975**	**0.967**	**0.964**	34.44

In Table 4.3, the proposed method is compared to other state-of-the-art methods for the pose dataset. I use accuracy, precision, recall, and F1 as my evaluation measures. I also compared the computation time in Table 4.3. As shown in Table 4.3, my methods outperform the rest by 10.2%, to 21.4% in terms of F1 score, which confirming that the estimated depth generation stage enhances the recognition accuracy. For processing time, though the proposed depth-estimation-based method takes twice as long as the conventional RGB-based method [3] and Skelton-based method [22], the proposed method can still perform posture recognition in real-time (about 29 fps) with higher recognition accuracy.

Figure 4.7 compares a color baseline with the proposed method, with the per-category for the pose database. The largest absolute gains are observed for pose2, pose3, pose6, and pose14. These are categories where depth information is of vital importance.

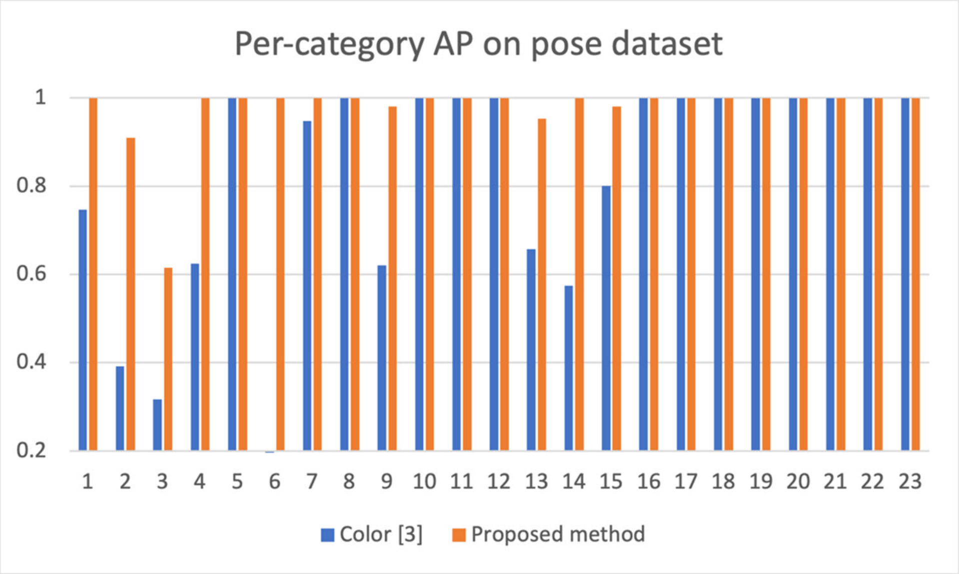


Figure 4.7: Per-category AP on the pose dataset: a color input baseline (blue) vs proposed method (orange).

4.3.2 Experiment on OUHANDS

4.3.2.1 OUHANDS Dataset

OUHANDS Dataset [20] contains 10 different hand gestures from 23 subjects and is split into training, validation, and testing sets with 1600, 400 and 1000 images, respectively. All sets come with corresponding segmentation masks, depth, and color images. The example of hand gestures is demonstrated in Figure 4.8.

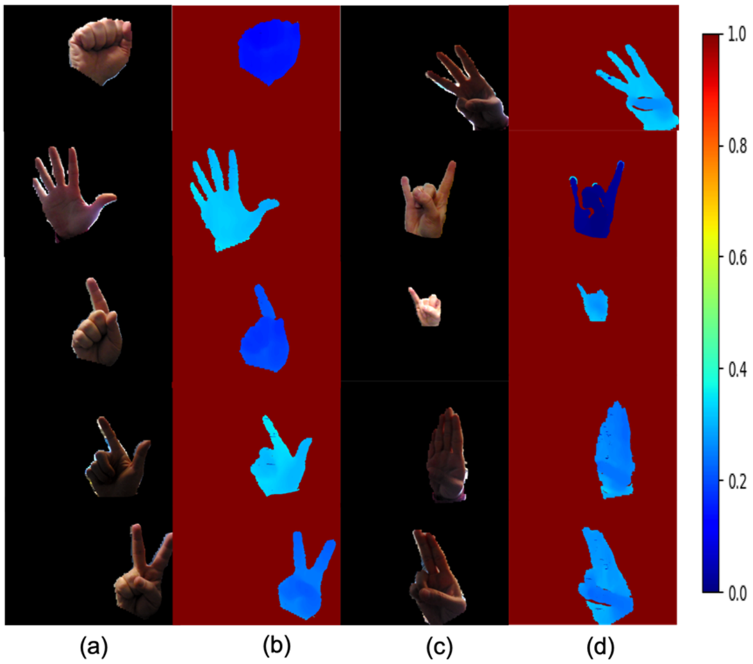


Figure 4.8: Samples from the OUHANDS train databases. (a)(c) columns show the hand region RGB data, while (b)(d) columns show the hand region depth data (with color bar)

4.3.2.2 Evaluation of the hand gesture recognition

Table 4.4 compares the performances of different network architectures on the OUHANDS dataset. The performance of the proposed method achieved the best accuracy when I applied my hybrid loss to coalesce the depth and color stream. The results emphasize the effectiveness of my fusion network architecture.

Table 4.4: Comparison of recognition accuracy on the OUHANDS test set.

Model	Acc	Precision	Recall	F1 score	Time (ms)
Baseline (ResNet 18)	0.888	0.890	0.888	0.887	23.6
Proposed method (estimated depth using hybrid loss stream only)	0.913	0.914	0.913	0.912	27.27
Proposed method two-stream late fusion (RGB, estimated depth using hybrid loss)	0.922	0.924	0.922	0.922	30.87

Figure 4.9 shows the category-wise comparison for color input baseline and proposed method on OUHANDs Dataset. The results show that the estimated depth information

improves the accuracy on most gestures, while reduces the accuracy on gestures 3, 8, and 9. The reason for accuracy reduction on gestures 3, 8, and 9 is because the estimation of the depth map for these gestures are not correct. The estimation means relative error for gestures 3, 8, 9 are 0.358, 0359, 0.452, respectively, while the mean relative error for the other 7 gestures is 0.34. Improvement of depth map estimation will be the future work.

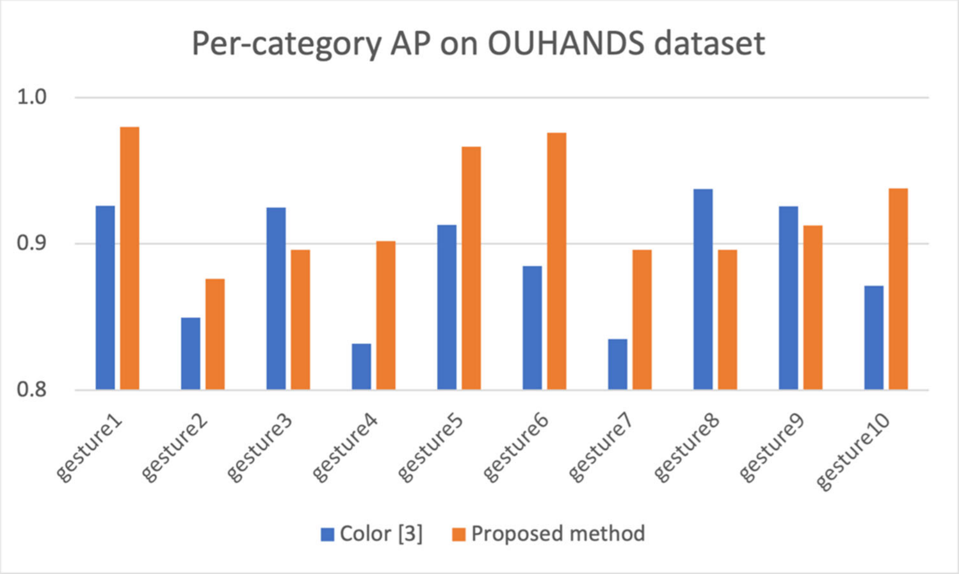
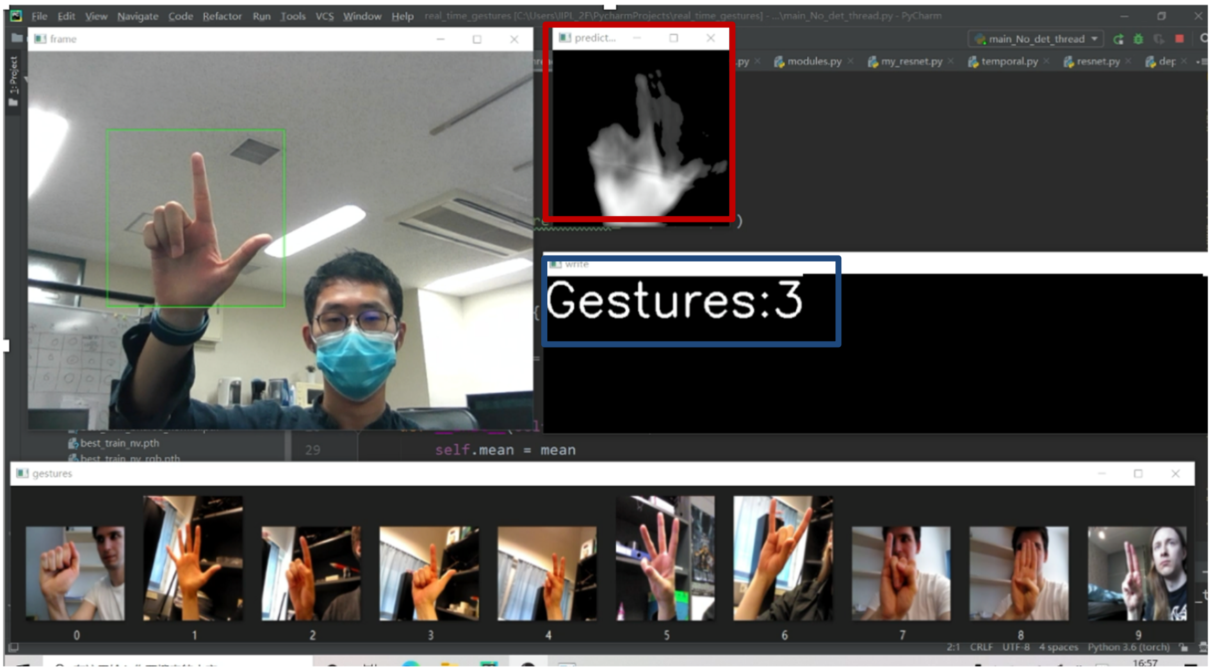


Figure 4.9: Per-category AP on the OUHANDs dataset: a color input baseline (blue) vs proposed method (orange).

4.3.2.3 Real-time gesture recognition system

I build a real-time gesture recognition system using the proposed method. The system is shown in figure 4.10. The green rectangle is the hand region color image, the red rectangle is the pseudo depth generated by my network. The real color image and generated depth image are used as input of the multimodal hand gesture network and get the recognition results (blue rectangle). The system can recognize hand gestures in real-time. In the future, high-accuracy touchless medical visualization can be realized only use a web camera in the operation room.


File Edit View Navigate Code Refactor Run Tools VCS Window Help
real_time_gestures [C:\Users\IIPL_2F\PycharmProjects\real_time_gestures] - ...\main_No_det_thread.py - PyCharm
frame
predict...
main_No_det_thread
modules.py
my_resnet.py
temporal.py
resnet.py
write
best_train_nv.pth
gestures
0
1
2
3
4
5
6
7
8
9
Python 3.6 (torch)

```
Gestures:3
```

```
self.mean = mean
```

Figure 4. 10: The real-time hand gesture recognition system. Green rectangle: color hand image (input); red rectangle: generated pseudo depth hand image; blue rectangle: classification results.

4.4 Chapter Summary

This Chapter proposed an RGB posture-recognition network based on a two-stage CNN architecture. To improve the recognition performance from color images, I generated an estimated depth posture image by a hybrid loss function incorporated in the generation module. The loss function captures the high-level features and recovers the sharp depth discontinuities. The proposed method was evaluated on the novel dataset of color-depth pose images and the public OUHANDS hand gesture dataset. The hybrid loss effectively and accurately generated depth posture images and the estimated depth image improved the accuracy of posture recognition. I am going to increase the human pose dataset including the back images of the participants and perform experiments to identify whether the participant is facing the camera or facing away from the camera.

Bibliography

1. Sajjad, F., Ahmed, A. F., and Ahmed, M. A., "A Study on the Learning Based Human Pose Recognition," 2017 9th IEEE-GCC Conference and Exhibition (GCCCE), Manama, 2017, pp.1-8.
2. Priya, B.G., Arulselvi, M., "Deep Learning for Human Pose Classification using Multi View Dataset," International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.
3. Pinto, R.F., Borges, C.D.B., Almeida, A.M.A., Paula, I.C., "Static Hand Gesture Recognition Based on Convolutional Neural Networks", Journal of Electrical and Computer Engineering, vol. 2019, Article ID 4167890, 12 pages, 2019. https://doi.org/10.1155/2019/4167890
4. Nishi, K., and Miura, J., "Generation of human depth images with body part labels for complex human pose recognition," Pattern Recognition, vol. 71, pp. 402–413, 2017.
5. Wang, H.K., Zhou, F.X., Zhou, W.J., Chen, L., "Human Pose Recognition Based on Depth Image Multifeature Fusion," Complexity. 2018. 1-12. 10.1155/2018/6271348.
6. Wan, C.D., Probst, T., Gool, L.V., Yao, A., "Crossing Nets: Combining GANs and VAEs With a Shared Latent Space for Hand Pose Estimation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 680-689
7. He, W., Xie, Z., Li, Y., Wang, X., and Cai, W., "Synthesizing Depth Hand Images with GANs and Style Transfer for Hand Pose Estimation," Sensors, vol. 19, no. 13, p. 2919, Jul. 2019.
8. Liu, J.Q., Furusawa, K., Tateyama, T., Iwamoto. Y., and Chen, Y.W., "An Improved Hand Gesture Recognition with Two-Stage Convolution Neural Networks Using a Hand Color Image and its Pseudo-Depth Image," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 375-379, doi: 10.1109/ICIP.2019.8802970.
9. Dadashzadeh, A.,Targhi, A.T., Tahmasbi, M., Mirmehdi, M., "HGR-Net: a fusion network for hand gesture segmentation and recognition." IET Comput. Vis. 13 (2019): 700-707.

10. Goodfellow, I. Pouget-Abadie, J. et al. "Generative adversarial nets," International Conference on Neural Information Processing Systems MIT Press, pp. 2672-2680, 2014.

11. Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A.A.," Image-to-image translation with conditional adversarial networks." In CVPR, 2017.

12. Gauthier. J., Conditional generative adversarial nets for convolutional face generation. Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester, 2014

13. Johnson, J., Alahi, A., and Li F.F., "Perceptual losses for real-time style transfer and super-resolution," in European Conference on Computer Vision, 2016.

14. Zheng, L., Tai, D., Forrester, C., Richard, T., Noah, S., Ce, L., Willium, T., "Learning the Depths of Moving People by Watching Frozen People," in CVPR 2019.

15. Kumari, S., Jha, R. R., Bhavsar, A., and Nigam, A.,"AUTODEPTH: Single Image Depth Map Estimation via Residual CNN Encoder-Decoder and Stacked Hourglass," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 340-344, doi: 10.1109/ICIP.2019.8803006.

16. He, K., Zhang, X.Y., Ren, S.Q., Sun. J., "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

17. Eigen, D., Puhrsch, C., and Fergus, R., "Depth map prediction from a single image using a multi-scale deep network." arXiv preprint arXiv:1406.2283, 2014. 2.

18. Krizhevsky, A., Sutsjever, I., and Hinton, G. "ImageNet classification with deep convolutional neural networks," Adv. Neural Inf. Process.Sys., pp9, 2012.

19. Microsoft, Kinect V2: https://developer.microsoft.com/en-us/windows/kinect/, Last Access: 2020.8.15.

20. Matilainen, M., Sangi, P., Holappa, J., Silvén, O."OUHANDS database for hand detection and pose recognition," 1-5. 10.1109/IPTA.2016.7821025.

21. Kingma, D.P., and Ba, J.,"Adam: A method for stochastic optimization." arXiv: 1412.6980, 2014.

22. Thomas, N., Kipf, M. Welling, Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR), 2017.

Chapter 5

Multimodal Deep Learning for Detection of Depressive Severity

5.1 Introduction

Depressive severity is widespread in the population and can negatively impact people's daily life in several ways. University students are at high risk of depressive severity as they can face intense academic, financial, and interpersonal pressures [1] while going through a critical period of transition from adolescence to adulthood and making many important life decisions [2]. Students with depressive severity can exhibit typical symptoms, such as low mood, loss of interest, and decreased energy. Such symptoms are a serious issue, and are especially significant for university students, since they can affect academic performance and health, and may in extreme cases lead to suicide [1].

People with depressive severity are screened using self-assessment questionnaires, such as Beck's Depression Inventory (BDI) [3] and the Centre for Epidemiologic Studies Depression Scale (CES-D) [4]. In this study, I define university students with depressive severity as those whose BDI-II and CES-D scores meet or exceed the depression assessment criteria but do not meet the diagnostic criteria for major depressive disorder given in the DSM-5 [5].
Depression, given its high incidence and negative impacts, such as impaired personal functions and social-economic burden [6, 7, 8], has become a serious social problem, worthy of the increased attention. Currently, the depression rate among Chinese university students has risen to 23.8% [2]. Previous research has shown that depressive severity experienced by the young is likely to persist into adulthood and develop into depressive disorder [9, 10]. Effectively recognizing such symptoms in university students can therefore help university mental health workers to identify and help them earlier, reducing the risk of depression.

Most existing studies into developing automated depression diagnosis systems have attempted to extract suitable features from a clinical interview dataset (e.g., the AVEC depression dataset) [11-13], focusing their analysis on patients with clinical depression from the western culture [11-15]. However, there has been little work on combining expression, action, and speech data in order to extract multimodal features and, in particular, there is currently no multimodal dataset based on Chinese university students with depressive severity.

The remainder of this Chapter is as follows. Section 5.2 introduces the related work including existing public dataset, depression detection using single modality and multi-modality. In Section 5.3 I described our private multimodal behavioral dataset of depressive symptoms (MB-DD), extractions of their audio and visual features and some experiments using deep learning on MB-DD. In Section 5.4, I presented a multi-modal adaptive fusion transformer network for depression detection using multi-task representation learning, which achieved the best results on the public Audio/Visual Emotion Challenge and Workshop (AVEC 2019) Dataset. I concluded in Section 5.5 with a summary.

Related Work

5.1.1 Public dataset for depression detection

Access to clinical data is extremely important for depression detection. Due to the sensitivity and privacy of clinical data, depression datasets are neither widely available for free. The current depression datasets are as follows: Black Dog Institute depression dataset (BlackDog) [64], DAIC-WOZ [66], Audio/Visual Emotion Challenge depression dataset (AVEC) [26], University of Pittsburgh depression dataset (Pitt) [65]. The Black Dog Institute is a clinical research facility in Australia, which collected a clinically depression dataset. The audio-video experimental process contains reading sentences and interviews. At the University of Pittsburgh (Pitt), a clinically depression dataset was collected during treatment sessions. A total of 57 depression were collected at seven-week intervals using HRSD clinical interview. DAIC-WOZ is partly available and be used as part of AVEC. The AVEC is the only fully public available for free download. AVEC 2019 is the ninth competition aimed at providing a common benchmark test set for multimodal information processing. Detecting depression with

AI is a sub-challenges of AVEC 2019. A summary of database for depression detection is shown in Table 5.1

Table 5.1: A summary of database for depression detection

Dataset	Subject	Modalities	Procedure	Depression Scale
AVEC2019 [26]	275	audio/video/Text	Human-computer interaction	PHQ-8 [53]
DAIC-WOZ [66]	189	audio /video/Text	Human-computer interaction	PHQ-8 [53]
BackDog [64]	80	audio /video	Watch clips, reading speech, structure interview	DSM-IV [67]
Pitt [65]	57	audio /video	HRSD clinical interview	DSM-IV [67]

5.1.2 Depression detection using single-modal information

As data used in Deep Learning for depression detection are time series, irrespective of how many modalities there are, it is important to effectively extract temporal information from every single modality. Currently, the most used methods for extracting temporal information for a single modality are RNN models, including LSTM and GRUs. For example, the baseline model of the AVEC 2019 DDS Challenge [26] used a single GRU layer to process time series to detect depression levels. [39] used a hierarchical Bi-LSTM to extract temporal information to obtain information with different temporal scales. [42] used the traditional LSTM structure to obtain sequential features for every single modality to estimate the levels of depression. Although RNN families are widely used for extracting temporal information, they still have some drawbacks, the most significant being the problem called Forgetting. The forgetting issue is explained as an RNN model that loses previous information when processing long-term sequences. Although LSTM and GRUs have been proposed to mitigate the negative impact of the forgetting problem, unsatisfactory results are achieved while processing extremely long-term sequences. This forgetting issue limits the sequential length that RNN models can process. The forgetting issue can be handled better now that the transformer model [41] has been proposed. As a transformer model [41] has a pure attention structure, the impact of forgetting is small, allowing the model to process longer sequences than traditional RNN families.

While original transformer models have been successfully used in natural language processing tasks, recent research studies have employed transformer models in other fields, such as image processing and emotion recognition image. In particular, [45] fused a CNN model and transformer model to process images, which has been called conformer. In the field of emotion recognition, [46] first used a transformer model to predict emotions. Because of the similarity between emotion recognition and depression detection, numerous research studies [47] have applied emotion methodologies to depression detection. In this work, I used a transformer model to predict the levels of depression; to the best of our knowledge, this is the first time a transformer model is used in this field.

5.1.3 Depression detection using multi-modal information

Multi-modal learning is one of the most important strategies in depression detection. As the data to be analysed in depression detection are composed of several modalities, such as video, audio, and text, it is relatively common to perform multi-modal learning. Currently, numerous research studies [38, 39] have proven that multi-modal learning can improve the accuracy and robustness of depression level prediction. The most used modalities include audio, videos, and texts, which are collected through interviews with patients suffering from depression, with their corresponding features, such as MFCCs and AUposes. For example, the AVEC 2019 DDS Challenge [26] dataset includes features extracted from original audios and videos, such as MFCC, eGeMAPS, and AUposes.

The multi-modal fusion strategy can be roughly divided into early fusion and late fusion. Early fusion means fusion of data at the feature level, whereas late fusion means fusion of data at the decision level. Nowadays, most methods fuse information in the early fusion stage. For instance, [48] used the bag-of-words model to encode audio and visual features and then fused them to perform multi-modal learning for depression detection [49] used texts generated from the original speech audio by Google Cloud's speech recognition service with their hidden embedding extracted from pretrained BERT [50] model while concatenating all modalities, achieving a concordance correlation coefficient (CCC) score of 0.69 on the AVEC 2019 DDS

Challenge dataset. Aside from audio, video, and text modalities, [7] employed body gestures as one of the modalities to perform early fusion. For late fusion, the most representative method is the baseline model of the AVEC 2019 DDS Challenge [26], which first obtains results from each unimodality and then takes the average as the final prediction.

However, most of the current methods did not explicitly weigh modalities with different performances, whether using early or late fusion. In my work, I propose an adaptive late-fusion strategy that can leverage the importance of different modalities. Specifically, I weight modalities according to their performances, which means that I assign high weights to modalities with high performance and low weights to those with poor performance to obtain final late-fusion results. According to the experimental results, we can infer that the proposed Adaptive Late-Fusion can improve the performance of depression detection.

5.2 Multimodal Behavioral Dataset of Depressive Symptoms (MB-DD)

In this section, I first described our private multimodal behavioral dataset of depressive symptoms (MB-DD), which is constructed under the collaboration with Prof. Huang Xinyin's Lab in Soochow University, China. Then I represented extraction of their audio and visual features and some experiments using deep learning on MB-DD. Figure 5.1 gives an overview of this study. First, the multimodal dataset (MB-DD) is created to investigate the relationship between university students' depressive severity and their observed behavior during several behavioral experiments. The dataset comprises two components: the behavioral dataset and the screening survey results. Later (in Section 5.3.3), I will extract visual audio features from part of these data to use them to construct a model (or mapping function) to investigate the relationship between participants' behavior and their depressive severity. In this study, I use the results of screening surveys as ground truth regarding depressive severity.

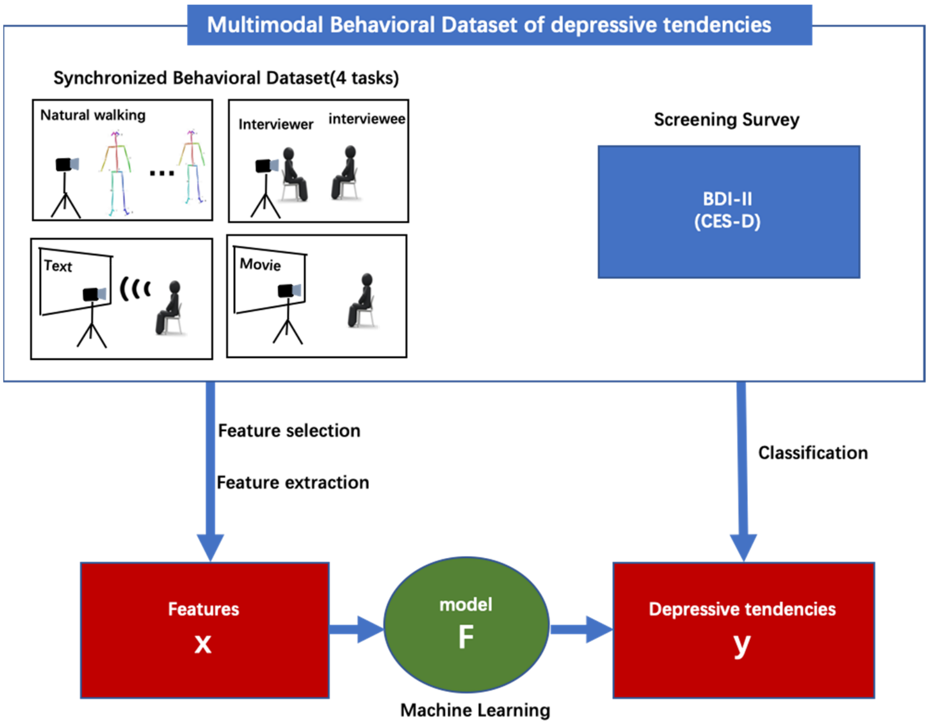


Figure 5.1: Overview of the study into the relationship between university students' depressive severity and their observed behavior.

5.2.1 Collecting Survey Data

This study was reviewed and approved by Soochow University in China. In this study, I used BDI-II screening survey data as ground truth regarding depressive tendencies. I used two scales (BDI-II and CES-D) to increase data credibility and eliminate participants whose scores differed significantly.

5.2.1.1 Beck Depression Inventory-II

The BDI is a 21-item self-reported depression metric. Each item is rated on a Likert scale with four possible answers, increasing in intensity from 0 to 3, yielding a total BDI score of between 0 and 63. In this study, I used the second BDI version, revised by Wang et al. [16]. There are four specific levels of the severity of depressive severity: 0 to 13 as minimal (no depression), 14 to 19 as mild, 20 to 28 as moderate, and 29 to 63 as severe [17]. For 2-class classification, 14 is the classification boundary. In this study, the BDI-II data's internal consistency was 0.88.

5.2.1.2 Centre for Epidemiologic Studies Depression Scale

The CES-D is a 20-item self-reported depression metric. Each item is rated on a Likert scale with four possible answers, increasing in intensity from 0 to 3, yielding a total CES-D score of between 0 and 60. The Chinese version of CES-D [18] was adopted in this study. I set the threshold for possible depression to 16, following the original author's recommendation [19]. In this study, the CES-D data's internal consistency was 0.86.

Participants were recruited by distribution and collection of questionnaires on campus. Students who met the screening criteria were invited to participate in the study by phone or text message. All participants were first taken through a consent process. They were then invited to complete the BDI-II and CES-D again, and the resulting scores were used to select participants for further experimental analysis.

102 participants (Chinese university students) were recruited for the study. The participants were divided into two groups: depressive persons (DP) and healthy persons (HP), according to their scores on standardized self-report questionnaires (BDI-II [19] and CES-D [20]). The DP group included 51 participants (26 males, 25 females): BDI-II ≥14 and CES-D ≥16, none of whom met the DSM-5 diagnostic criteria for major depressive disorder. The HP group included 51 participants (26 males, 25 females): BDI-II < 14 and CES-D < 16, none with histories of mental illness. There was no age difference between the DP and HP groups (t (100) =0 .80, p = 0.43). The BDI-II (t (100) = 14.38, p < 0.001) and CES-D (t (100) =14.17, p < 0.001) scores were significantly higher in DP than in HP. Differences in the groups' demographic and psychological characteristics were presented in Table 5.2. A preliminary study of this database is referenced in [21].

Table 5.2: Differences in the groups' demographic and psychological characteristics

	DP **(*N=51*)**	**HP** **(*N=51*)**
Age(*M*±*SD*)	18.98±0.91	18.84±0.83
Gender (*n*)		
Male	26	26
Female	25	25
BDI-II(*M*±*SD*)	21.31±6.72	5.45±4.12
CES-D(*M*±*SD*)	24.18±6.82	7.53± 4.89

5.2.2 Acquiring Behavioral Data

Next, four experimental tasks were tried out for data collection, and the data acquisition system was shown in Figure 5.2. The four experimental tasks in this study were designed based on preliminary experiments with reference to relevant studies [22, 23, 24].

In this subsection, I will introduce the multimodal data acquisition system used to build the behavioral dataset. Participants sat 2.7m away from the display screen, which was 1.9m×1.06m. A web camera (Logitech C920) was set up directly 1.2m away in front of them to synchronously collect their expression and voice information at a resolution of 1920 × 1080 with a frame rate of about 50 frames per second.

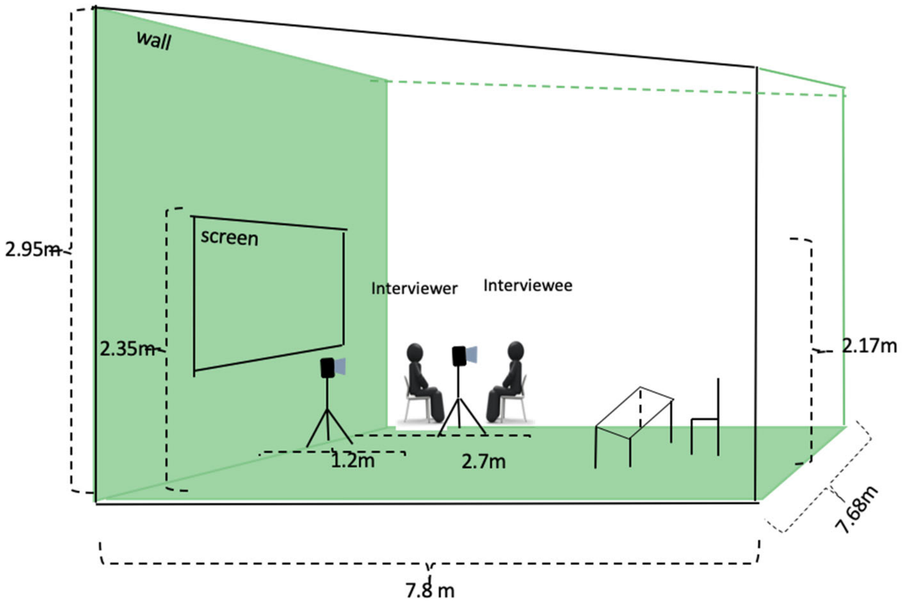


Figure5.2: Illustration of the data acquisition system.

As shown in Figure 5.1, the experimental tasks in the behavioral database in this study included four tasks: natural walking, natural situational interview, reading emotional text and freely watching emotional videos. The four experimental tasks were completed on the same day, and each participant completed all the experimental tasks in the order of Task 1, Task 2, Task 3 and Task 4 (the sequence arrangement of the four experimental tasks was adjusted and determined according to the feedback of the subjects in the pre-experiment and the coherence of the whole experiment). Adequate rest time was set between tasks to reduce the interference between different tasks.

In task 2, I designed 13 questions based on the diagnostic criteria for major depressive disorder given in the DSM-5 [5] and the Hamilton Depression Rating Scale [20]. These questions were designed to elicit spontaneous speech from the participants, together with related facial expressions and actions. During this process, I also wanted to ensure that the participants were not clinically depressed. Those who answered yes to fewer than five of the first nine questions were not asked the remaining ones (10–13). Table 5.3 list 13 main topics covered during the interviews (task 2).

To facilitate the follow-up research to explore the cross-valence stability of the interview questions, participants would be asked three types of emotional questions at the beginning of the interview: (1) Neutral question: Can you tell me something about your recent study and life? (2) Positive question: Please share with me a good memory and describe the scene at that time. (3) Negative question: Please share with me a sad memory and describe the scene at that time. The list of topics is listed in Table 5.3.

Table 5.3. List of topics covered during the interviews (task 2).

Topic	Sample Questions
1	How has your mood been for the last two weeks? -Have you felt sad for most of the days?
2	What are you usually interested in? -Have you been interested in this during the last two weeks? -Have these activities brought you pleasure during this time? -Has your interest in other topics diminished?
3	Has your appetite changed at all during the last two weeks? -Has your weight changed during this time? -By how much has it increased or decreased? - Has it changed by more than 5% of your original body weight?
4	How have you slept during the last two weeks? -Did you have insomnia (such as having difficulty falling asleep, waking during the night, waking in early hours and unable to fall asleep again) or sleep too much?
5	Here, notes were made of the behavior observed during the interview, such as fidgetiness, playing with hands, hair, inability to sit still, standing during the interview, hand wringing, nail biting, hair pulling, biting of lips.
6	How has your energy been over the last two weeks? -Have you always been tired? - Have you experienced back pain, headaches, or muscle pain, or heaviness in your limbs, head, or back?
7	Have you blamed yourself for anything over the last two weeks? -Have you felt guilty for most of the day during the last two weeks? -Have you felt worthless during this time? - Did it last for most of the day during the last two weeks?
8	Have you felt unable to think over the last two weeks? -Did this last for most of the day during the last two weeks? -Have you been able to concentrate on what you were doing during this time? - Did it last for most of the day during the last two weeks? -Have you felt hesitant to do

	something during this time? -Did it last for most of the day during the last two weeks?
9	Have you experienced any extreme thoughts or behaviors over the last two weeks, such as hurting yourself or committing suicide? -Did you act on them?
10	Have the problems you've talked about had a negative impact on your social life, studies, or daily life, giving you pain or discomfort?
11	Are these problems related to a particular substance or disease?
12	Have you ever had any psychiatric disorders (schizophrenia spectrum disorders or other psychiatric disorders)? -Were/was these/this associated with the onset of the problems you've talked about?
13	Have you had a remarkably persistent high level of emotional ego-inflation or mood irritability, or an unusually persistent increase in activity or energy most of the day more than 4 days a week?

Task 3 was inspired by related work [25] to collect more audio information from the subjects. The emotional sentences are listed in Table 5.4.

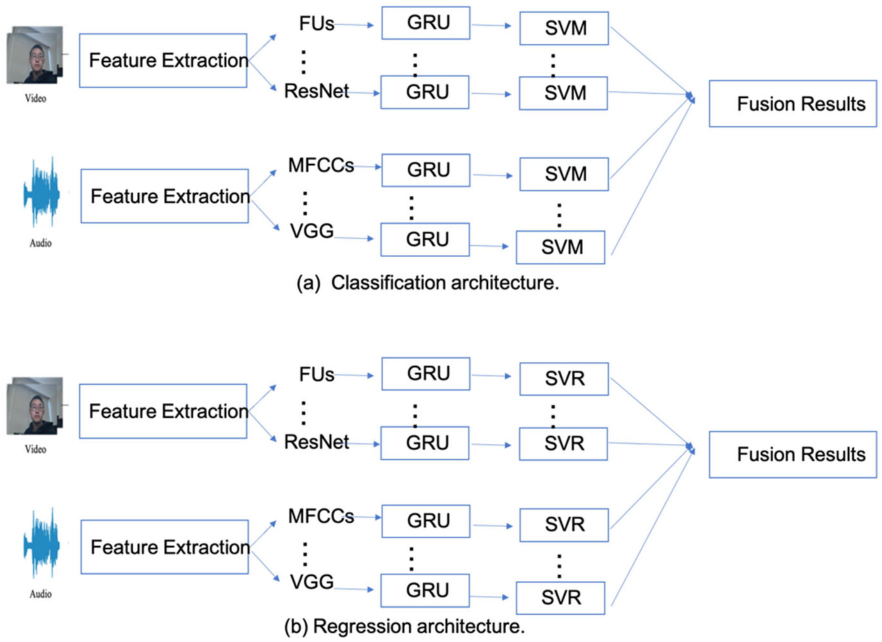
Table 5.4. List of emotional sentences. *n* is the number of key frames in each sentence (task 3).

Emotional Type	Sentence ID	Content written in Chinese (translate to English)	*n*
Positive	No.1	盼啊！盼啊！眼看春节就快到了。 (Wish ah! Wish ah! The Spring Festival is coming soon.)	12
	No.2	想到这，我不由得笑了起来。 (Thinking about it, I couldn't help laughing.)	11
	No.3	在春节前，人们个个喜气洋洋，个个精神饱满。 (Before the Spring Festival, people are all beaming, and in high spirt.)	18
	No.4	逛街的人络绎不绝，有的在买年画，有的在买年货。 (People go shopping in an endless stream, some are buying New Year pictures, some are buying New Year goods.)	20
	No.5	有的围着火炉看电视，还有的人在打麻将打扑克等等，不一而足。 (Some were watching TV by the fire, others were playing mah-jongg and poker, and so on.)	26
	No.6	大年三十，人们常常玩到深夜，嘴里啃着美味水果，手里燃放烟花爆竹。 (On New Year's Eve, people often play late into the night, eating delicious fruit and setting off fireworks in their hands.)	28
	No.7	大人小孩都载歌载舞，忘情地玩个痛快。 (Adults and children are singing and dancing and enjoy themselves.)	16
Neutral	No.1	卢沟桥位于北京广安门外永定河上，距天安门 15 千米 (Lugou Bridge is located on the Yongding River outside Guang'anmen Square in Beijing, 15 kilometers away from Tiananmen Square.)	23
	No.2	它始建于金代大定年间，历时 3 年建成，定名为“广利桥” (It was built in the Dading period of the Jin Dynasty, took 3 years to build, and was named "Guangli Bridge".)	22
	No.3	又因永定河旧称卢沟河，所以广利桥俗称卢沟桥。 (Because the Yongding River was formerly known as the Lugou River, the Guangli Bridge is commonly known as Lugou Bridge.)	20

	No.4	卢沟桥是北京地区现存的最古老的一座联拱石桥 (Lugou Bridge is the oldest existing multi-arch stone bridge in Beijing.)	21
	No.5	明清两代都有重修，现在所见到的为 1986 年重修复原后的石桥 (The stone bridge was rebuilt in the Ming and Qing dynasties, and what we see now is the stone bridge that was rebuilt in 1986.)	28
	No.6	桥长 266.5 米，桥面宽 9.3 米，为花岗岩所修成。 (The bridge is 266.5 meters long and 9.3 meters wide. It is made of granite.)	24
Negative	No.1	三十三年前的一次车祸让妈妈永远的离开了我们 (Thirty-three years ago, a car accident took my mother from us forever.)	21
	No.2	妈妈您在另一个世界过得还好吗 (Mom, how is life with you in the other world?)	14
	No.3	儿子好想念您呀 (Your son misses you so much!)	7
	No.4	这么多年来，儿子一时一刻没有忘记您那慈祥的笑容 (Over these years, your son never forgot your kind smile for a moment.)	22
	No.5	虽然您离开时我只有十三岁，虽然生前一张照片也没留下，可是儿子永远也忘不了您 (Although I was only thirteen when you left me, and you had not left a single picture yet, I will never forget you.)	35
	No.6	多少次夜里梦见您的身影：多少次梦中想您哭醒 (I have dreamed of you many times during the night and have cried for you many times in my dreams.)	20
	No.7	妈妈，您怎么那么狠心就扔下我们不管了呢？ (Mom, why did you leave us so cruelly？)	18

5.2.3 Baseline Estimation

In this subsection, I make a preliminary application of the multimodal dataset established in the previous section, to evaluate the feasibility of this dataset in predicting university students' depressive severity using data from Task3 as an example. Figure 5.3 shows the architecture of the proposed model. The deep neural network model consists of three parts: (1) the subnetworks for each single modality feature extraction. (2) the gated recurrent unit (GRU) network for each audiovisual representation. (3) the final decision layer that detection depressive severity.



(a) Classification architecture.

(b) Regression architecture.

Figure 5.3: The architecture of the proposed model. The unimodal features are extracted separately and concatenated in a decision strategy. Classification architecture (a) and regression architecture (b).

5.2.3.1 Feature Extraction

In this subsection, I will explain how I calculated the baseline set of behavioral features, which can be used to investigate the relationship between university students' depressive severity and their observed behavior. The feature sets are inspired form AVEC 2019 [26]. For ethical reasons, I have not published raw video. The numbers of all features of audio and video are summarized in Table 5.5.

Table 5.5. Numbers of features of audio and video.

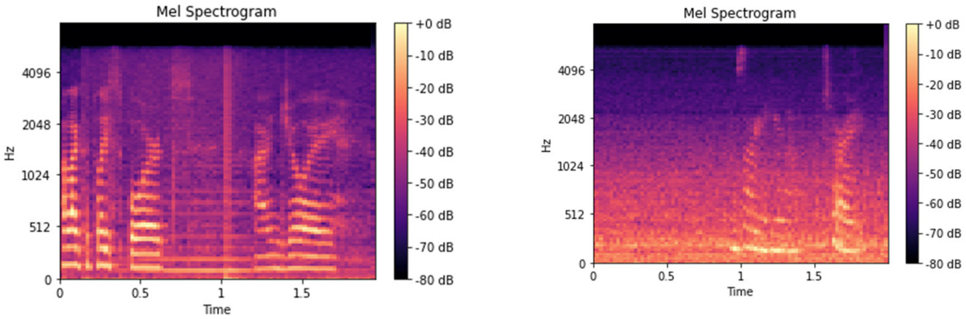
	Low-level feature				Middle-level features (BoW)		High-level features	
	MFCCs	eGeMAPS	MFCCs-F	eGeMAPS-F	BoW-m	BoW-e	DNet	VGG
Audio	39	88	78	176	100	100	1920	4096
	Low-level feature				High-level features			
	FAUs				ResNet (ImageNet)	ResNet (Affwild)		VGG (Affwild)
Video	17				2048	2048		4096

Audio Feature

The first step in analyzing the prosodic features of a person's speech is to isolate it from silence, other speakers, and noise. As audio features, I use the openSMILE [27] toolkit to compute the low-level features. To calculate functionals, several statistical measures are applied to normalize the low-level features.

For middle-level feature, the bags-of-words (BoW) model, which originates from text processing, represents the distribution of LLDs according to a dictionary learned from them is used. The bags-of-words (BoW) involves four major steps: (1) local descriptors detection from traing data; (2) codebook learning using the local descriptors; (3) coding the local descriptor in terms of the learning codebook; (4) pooling operation by accumulating the codes of local descriptors of the LLDs into a fix-length representation feature. The open-source toolkit oepnXBOW [29].

For deep representations, inspired by the development of deep learning in image processing, spectrogram images of speech instances are fed into pre-trained image recognition CNNs using VGG-16 [30] and DENSENET-201 [31] to extract high-level features. The spectrogram is calculated by the short-time Fourier transform on overlapping windowed segments of the signal. The mel spectrogram is a spectrogram where the frequencies are converted to the mel scale. The mel scale is a perceptual scale of pitch which equal distances in pitch sounded equally distant to the listener. In particular, the audio waves are first transformed into mel spectrogram images with 80 mel-frequency bands with 4s window width and a hop size of 1 s. Figure 5.4 shows examples of mel spectrogram for depression and non-depression participants.



(a)

(b)

Figure 5.4: Examples of mel spectrogram for (a) depression participant and (b) non-depression participant.

The audio features are extracted as follows:

1. **Mel Frequency Cepstral Coefficients (MFCCs),** which are a compact representation of the short-time power spectrum of speech. Figure 5.5. shows the process of creating MFCC features.

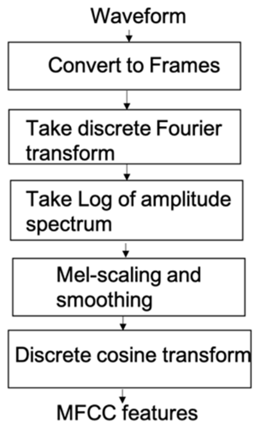

Waveform
Convert to Frames
Take discrete Fourier transform
Take Log of amplitude spectrum
Mel-scaling and smoothing
Discrete cosine transform
MFCC features

Figure 5.5: Process to create MFCC features.

2. **Extended Geneva Minimalistic Acoustic Parameter set (eGeMAPS)** contains 88 dimensions features, resulting in a feature vector with a dimension of 88. The minimalistic acoustic parameter set contains the following compact set of 18 Low-level descriptors (LLD), sorted by parameter groups. Some typical features of eGeMAPS are shown in Table 5.6.

Table 5.6: Some typical features of eGeMAPS.

Frequency related parameters	
Pitch	logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0)
Jitter	deviations in individual consecutive F0 period lengths
Formant 1, 2, and 3 frequencies	centre frequency of first, second, and third formant
Formant 1	bandwidth of first formant
Energy/Amplitude related parameters	
Shimmer	difference of the peak amplitudes of consecutive *F0* periods
Loudness	estimate of perceived signal intensity from an auditory spectrum
Harmonics-to-Noise Ratio (HNR)	relation of energy in harmonic components to energy in noise like components
Spectral (balance) parameters	
Alpha Ratio	ratio of the summed energy from 50–1000 Hz and 1–5 kHz
Hammarberg Index	ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region
Spectral Slope 0–500 Hz and 500–1500 Hz	linear regression slope of the logarithmic power spectrum within the two given bands.
Formant 1, 2, and 3 relative energy,	as well as the ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F0.
Harmonic difference H1–H2	ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2).
Harmonic difference H1–A3	ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3).

3. **MFCCs-F** represents for the functionals of MFCCs. The low-level feature MFCCs is summarized over time by computing their mean and standard deviation using a sliding window of 4 s length, and a hop size of 1 s.
4. **EGEMAPS-F** represents for the functionals of eGeMAPS. The eGeMAPS features is is summarized over time by computing their mean and standard deviation using a sliding window of 4 s length, and a hop size of 1 s.
5. **BoW-m** represents the bags-of-words representation of MFCCs feature. The codebook size is 100. MFCCs feature is processed a summarized over a block of a 4 s length duration.
6. **BoW-e** represents the bags-of-words representation of eGeMAPS feature. The codebook size is 100. eGeMAPS feature is processed a summarized over a block of a 4 s length duration.

7. **DNet** represents the deep representation using pre-trained DenseNet-201. The input is the mel spectrogram image. A 1920-dimnesional feature vector is extracted from the last average pooling layer of DenseNet-201.
8. **VGG** represents the deep representation using pre-trained VGG-16. The input is the mel spectrogram image. A 4096-dimnesional feature vector is extracted from the second fully connected layer in VGG-16 networks.

Visual Feature

For low-level descriptors of visual features, I use the OPEN-FACE toolkit [33] to extract the intensities of 17 facial action units (FAUs) for each video frame (Figure 5.6), along with a confidence measure.

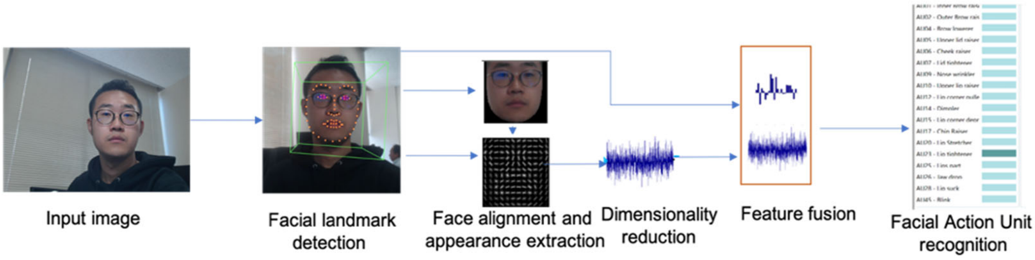

Input image
Facial landmark detection
Face alignment and appearance extraction
Dimensionality reduction
Feature fusion
Facial Action Unit recognition

Figure 5.6: Low-level descriptors extraction of visual features using OpenFace, including facial landmark detection, head pose and eye gaze estimation, facial action unit recognition.

For deep visual representations, I employed a VGG-16 [30] network and a ResNet-50 network that are pre-trained with the Affwild dataset [34], which focuses on human affect understanding. In particular, the OPEN-FACE toolkit [33] is used to detect the face region and subsequently performed face alignment. Following that, the aligned face images are forwarded through the two pre-trained models, respectively.

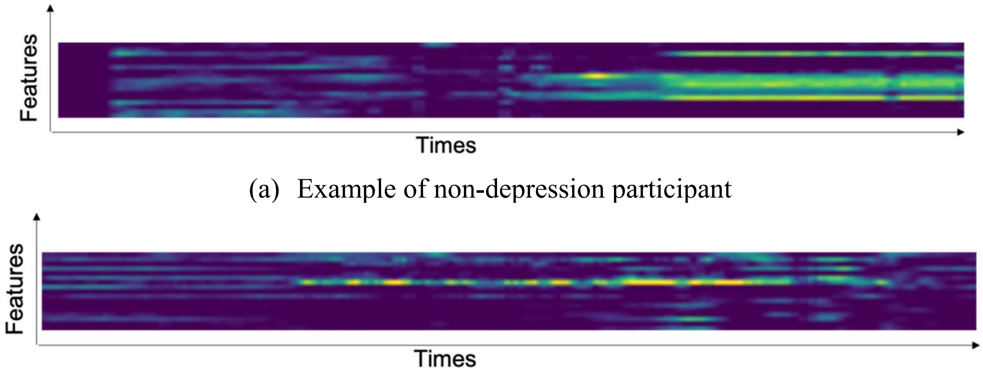
The visual features are extracted as follows:

1. **FAUs** represents facial action units. The description of action units is shown in Table 5. 7. Examples of visualization of FAUs features generated from a clip of video for depression and non-depression participants are shown in Figure 5.7.

Table 5.7: Description of action units.

1	Inner brow raise	10	Upper lip raiser	25	Lips part
2	Outer brow raise	12	Lip corner puller	26	Jaw drop
4	Brow lowerer	14	Dimpler	45	Blink
5	Upper lid raiser	15	Lip corner depressor		

6	Check raiser	17	Chin raiser		
7	Lid tightener	20	Lip strecher		
9	Nose wrinkler	23	Lip tightener		

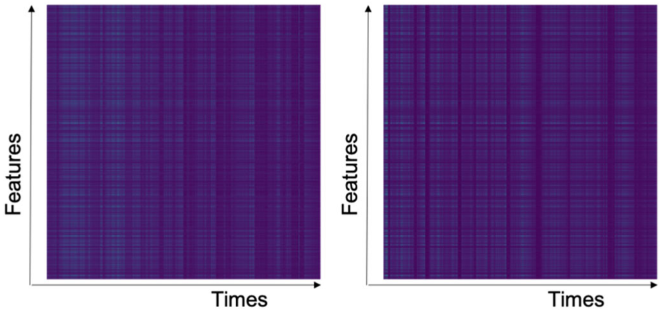


(a) Example of non-depression participant

(b) Example of depression participant

Figure 5.7: Visualization of FAUs features generated from a clip of video.

2. **ResNet（ImageNet)** represents the representation using ResNet-50 pretrained by ImageNet. A 2048-dimensional deep feature vector from ResNet are extracted for each frame. Figures 5. 9. Shows the examples of visualization of ResNet features generated from a clip of video for depression and non-depression participants.



(a) Non-depression participant (b) Depression participant

Figure 5.9: Visualization of ResNet features generated from a clip of video.

3. **ResNet (Affwild)** represents the representation using ResNet-50 network pretrained by Affwild database. A 2048-dimensional deep feature vector from ResNet-50 network are extracted for each frame.
4. **VGG (Affwild)** represents the representation using VGG pretrained by Affwild database. A 4096-dimensional deep feature vector from VGG network are extracted for each frame.

5.2.3.2 Detection Model

Finally, I will introduce the baseline gated recurrent unit (GRU) network and a late fusion strategy to combine audio and visual modalities in this subsection.

A gated recurrent unit (GRU) was proposed by Cho et al. [35] to make each recurrent unit adaptively capture dependencies of different time scales. As well as the LSTM unit, the GRU has gating units that modulate the flow of information inside the unit. However, without having any separate memory cell [36]. I use a gated recurrent unit (GRU) network with two-layers, each having 64 nodes for their hidden layers, for each audio-visual feature. The GRU is then followed by a fully connected neural network that has one hidden layer with 32 nodes, followed by a single linear layer to map to the desired output size of one for depressive severity regression task and output of four for the classification task.

I define the two tasks used in the experiments: the depressive severity regression task and the depressive severity classification task. I used SVM for classification task and SVR for regression task as shown in Figure 5.3 (a) and (b), respectively. In the depressive severity regression task, I predict its BDI-II score, which range from 0 to 63 in the database. The loss function for depressive severity regression task is concordance correlation coefficient (CCC) Loss function. The CCC loss function (L_{ccc}) can be defined as Equation (5.1) to maximize the agreement between true value (y) and prediction depressive symptoms degree ($\hat{y}$).

$$L_{ccc} = 1 - \frac{2p_{\hat{y}y}\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_{\hat{y}}^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{5.1}$$

where $p_{\hat{y}y}$ is the Pearson coefficient correlation between $\hat{y}$ and y, σ is standard deviation, and μ is a mean value.

In classification task, I discretize the BDI-II score into 4 classes [17]: minimal (no depression) [0-13], mild [14-19], moderate [20-28], and severe [29-63]. I treat this problem as multi-class classification and cross-entropy loss is used. The cross-entropy loss can be defined as:

$$L_{CE} = -\sum_{i}^{n} y_i \log(p_i) \tag{5.2}$$

where y_i it the truth label and p_i is the SoftMax probability for the i^{th} class.

Late fusion is the most used method for multimodal depression severity detection. I train a classifier for each modality and merges the decision values from each unimodal modal into a unified decision using averaging sum.

As shown in Figure 5.3, the model has been obtained as follows. First, the single feature streams are trained separately using the ground truth. The output of the 8 audio single streams and 4 visual streams are used as inputs to the decision fusion.

5.2.4 Experiment Results

In this section, I report the results of my model variants described in Section 5.2.3.

5.2.4.1 Experimental Setup

The novel multimodal behavioral dataset of depressive severity for task 3, which is described in Sec.5.3.2, is used in experiments. There are 102 participants in the dataset. I divide them into training, development, and test sets. The distribution of the training, development and test splits is summarized in Table 5.8.

Table 5.8: Distribution of the training, development, and test splits.

Task		Train	Dev	Test
Regression Task		72	10	20
Classification Task	Minimal [0-13]	36	5	10
	Mild [14-19]	14	2	4
	Moderate [20-28]	16	2	4
	Severe [29-38]	6	1	2

For evaluation on the test set, I use the best performance model on the development set. To handle the bias, I converted the BDI-II score labels to floating point numbers by downscaling with a factor of 38 prior to train. The RMSE results are reported using the original BDI-II scale. The model is implemented using a PYTORCH framework and is trained with an ADAM optimizer.

5.2.4.2 Evaluation Functions

To evaluate regression/classification results, I use well-known evaluation metrics that are standard for depressive severity detection.

I use the concordance correlation coefficient (CCC) as a measure of estimated scores (regression task), which is the common metric in dimensional depressive severity detection to measure the agreement between true depressive severity degree (y) with predicted depressive severity $(\hat{y})$. The benefits of using CCC are not biased by changes in scale and location, and elegantly includes information on both precision and accuracy in a single evaluation measure. The CCC is formulated as Equation (5.3):

$$CCC = \frac{2p_{\hat{y}y}\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \qquad (5.3)$$

where $p_{\hat{y}y}$ is the Pearson coefficient correlation between $\hat{y}$ and y, σ is standard deviation, and μ is a mean value. This CCC is based on Lin's calculation [37]. The range of CCC is from -1 to 1, which -1 perfect disagreement and 1 perfect agreement.

I also use the Root Mean Squared Error (RMSE), which is defined as Equation (5.4), as another measure for regression task.

$$\text{RMSE} = \sqrt{\frac{\sum_i^N (y_i - \hat{y}_i)}{N}} \qquad (5.4)$$

For classification task, the accuracy (denoted as *Acc*) is defined on all test samples and it the fraction of predictions that the model got right, Total accuracy reaches its best value at 1 and its worst score at 0. It is defined as:

$$Acc = \frac{\text{Number of Correnc Predictions}}{Total\ Number\ of\ Predictions} \qquad (5.5)$$

5.2.4.3 Results

To demonstrate the effect of each feature, I summarized the results of using each audio feature and visual feature in Table 5.9 and 5.10, respectively. CCC and RMSR are results of the regression task and *Acc* is the results of classification task. The best result on each measure is highlighted in bold.

Table 5.9: The results of using each audio feature on development dataset and test dataset.

Negative

Partition		Audio							
		Low-level features				Middle-level Features (BoW)		High-level features (DL)	
		eGeMAPS	MFCCs	eGeMAPS-F	MFCCs-F	BoW-M	BoW-e	DNet	VGG
Dev	CCC	0.04	0.32	0.16	-0.14	**0.54**	0.10	0.26	0.31
	RMSE	23.07	10.59	12.77	17.00	9.90	10.48	11.25	**9.81**
	Acc	0.50	0.50	**0.70**	**0.70**	0.50	0.60	0.60	**0.70**
Test	CCC	-0.06	0.16	-0.03	**0.31**	-0.03	-0.19	0.20	-0.12
	RMSE	22.72	14.62	17.46	**11.37**	14.95	13.61	12.61	13.28
	Acc	0.10	0.25	0.30	**0.35**	0.25	0.30	0.25	0.25
Neutral									
Dev	CCC	0.43	0.12	**0.79**	0.51	0.17	0.22	0.51	0.49
	RMSE	10.69	15.35	**5.79**	9.99	14.91	10.31	9.47	9.47
	Acc	0.60	0.50	0.70	**0.80**	0.70	0.60	0.60	0.50
Test	CCC	-0.21	0.07	0.19	-0.15	**0.29**	0.15	0.18	0.10
	RMSE	14.31	13.26	11.95	16.22	12.29	**9.96**	13.18	11.68
	Acc	0.15	0.20	0.25	0.15	0.25	**0.30**	0.25	**0.30**
Positive									
Dev	CCC	0.36	0.24	**0.87**	0.43	0.57	0.39	0.60	0.47
	RMSE	10.94	12.47	**4.92**	11.80	9.24	9.38	8.65	9.67
	Acc	**0.60**	**0.60**	**0.60**	**0.60**	**0.60**	0.50	**0.60**	**0.60**
Test	CCC	0.16	-0.14	-0.50	**0.52**	0.27	-0.98	-0.23	0.05
	RMSE	12.32	17.50	19.23	**10.80**	12.53	12.82	13.75	12.67
	Acc	**0.35**	0.25	0.30	**0.35**	0.30	**0.35**	0.30	**0.35**

Table 5.10: The results of using each visual feature on development dataset and test dataset.

Negative					
Partition		Visual			
		Low-level features	High-level features (DL)		
		FAUs	ResNet (Affwild)	ResNet (ImageNet)	VGG (Affwild)
Dev	CCC	0.56	0.68	**0.72**	0.19
	RMSE	9.56	7.22	**7.02**	10.23
	Acc	**0.70**	0.60	0.60	0.50
Test	CCC	0.002	-0.06	-0.24	**0.008**
	RMSE	15.93	12.91	15.19	**11.42**
	Acc	**0.40**	0.25	0.30	0.30
Neutral					
Dev	CCC	0.65	0.47	**0.72**	0.44
	RMSE	8.33	9.20	**6.29**	8.83
	Acc	**0.70**	0.50	0.60	0.50
Test	CCC	-0.27	**-0.10**	-0.15	-0.12
	RMSE	19.79	16.84	13.74	**12.59**
	Acc	0.25	0.25	**0.30**	0.25
Positive					
Dev	CCC	0.40	0.78	**0.80**	0.39
	RMSE	9.73	6.93	**6.20**	10.04
	Acc	0.50	**0.60**	**0.60**	0.50
Test	CCC	**0.29**	-0.23	0.09	0.18
	RMSE	13.60	14.99	**10.77**	11.59
	Acc	**0.35**	0.25	0.2	0.30

With respect to regression task, the best results in terms of CCC score from audio features was achieved with BoW-M, EGEMAPS-F, EGEMAPS-F for negative, neutral, positive, respectively. And the model with Res-ImageNet features achieved the best result for visual features in all three valences. These results indicate the low-level features are more useful for audio, while representations learnt by deep neural networks are more powerful for visual. In term of valence, the positive-emotional speech achieved best results both in audio-based and visual-based depressive severity regression.

The results of fusing all the features (multi-modal features) are summarized in Table 5.11. Compared with Tables 5.9 and 5.10, one can conclude that the decision fusion outperforms any single modals, which indicates that the feature fusion may provide the complementary information for detection of depressive severity.

Table 5.11: The results of fusing all audio and visual features (multi-modal features) on development dataset and test dataset.

Negative		
Dev	CCC	0.52
	RMSE	8.38
	Acc	0.40
Test	CCC	-0.05
	RMSE	10.59
	Acc	0.45
Neutral		
Dev	CCC	0.61
	RMSE	7.22
	Acc	0.60
Test	CCC	-0.03
	RMSE	10.64
	Acc	0.55
Positive		
Dev	CCC	0.71
	RMSE	6.32
	Acc	0.40
Test	CCC	0.06
	RMSE	10.13
	Acc	0.50

5.3 Multimodal Adaptive Fusion Transformer Network for Detection of Depressive Severity with Public AVEC Dataset

5.3.1 Proposed Methods

In this subsection, I first show an overall description of my multi-modal adaptive fusion transformer network and then provide a detailed description of the transformer encoder module, encoding the time series data from each modality. Subsequently, I elaborate on how my multi-task methods use a multi-task representation learning network for PHQ-8 regression and 5-class classification. Finally, I fuse acoustic and visual modalities in Adaptive Late-Fusion to conduct the final depression level prediction. The architecture of the proposed method is presented in Figure 5.10.

To illustrate the effectiveness of the transformer model in depression detection, I employ the Transformer Encoder to extract temporal information. After features from every modality are processed by the *Transformer Encoder* and the *Embedding FC Block* presented in Figure 5.10, they are fed to two *FC Blocks* designed for multi-task learning, which will be later described in more detail in the Multi-Task Learning sub-section.

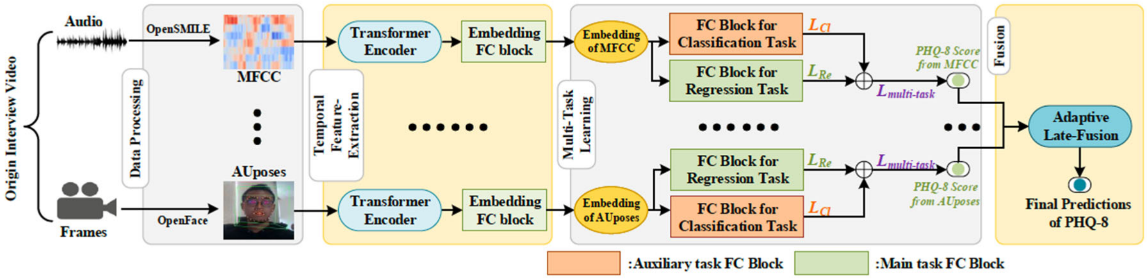


Figure 5.10: Overview of the proposed method. The origin data is firstly processed in the Data Processing Stage which has been done in the AVEC 2019 DDS dataset. Then the Transformer Encoder is used for extracting temporal information. The Embedding FC Block combined by a RELU activation layer, a dropout layer and a fully connected layer is used to extract the hidden embeddings representing each kind of feature. The embeddings from each kind of feature are fed to two FC blocks to perform multi-task predictions. Finally, the results from different features are fused in the Adaptive Late-Fusion to predict the final results. The L_{Re} means the Concordance Correlation Coefficient Loss for PHQ-8 regression and the L_{Cl} means the Cross-Entropy Loss for 5-class classification.

5.3.1.1 Input Stream

As data obtained from AVEC 2019 DDS [26] have been processed to specific features, the data pre-processing stage in Figure 5.11 can be skipped. In the AVEC 2019 DDS Challenge dataset, two main modalities can be obtained, namely, audio and video modalities. Each modality type contains several kinds of features, such as MFCC from audio and AUposes from video. For every type of feature obtained from the AVEC 2019 DDS Challenge dataset, the model is independently trained. The results obtained from each type of feature are fused in the independent stage called Adaptive Late-Fusion. For every type of feature, the transformer model is designed to extract temporal information, and its detailed structure will be described in the next section, i.e., Transformer Encoder. Suppose pre-processed features have the shape of R^{bs*t*d}, where bs standard is the batch size, t standard is the temporal frames, and d standard is the feature dimension. After they have been processed by the Transformer Encoder, the features with the shape of R^{bs*t*d} are averaged in the temporal t dimension to obtain the shape of R^{bs*d}. The averaged features are fed to the *Embedding FC Block* to obtain features with the same dimension, which are treated as hidden embeddings representing every feature from each modality. Each *FC Block* consists of a rectified linear unit (ReLU) activation layer, a dropout layer, and a linear layer. The dropout layer in the *FC Block* is designed to overcome overfitting during training. The hidden embeddings are finally passed to the two FC Layer Blocks to perform predictions on two tasks: PHQ-8 regression and 5-class classification. After the results from each modality's feature are obtained, I employ Adaptive Late-Fusion to obtain the final prediction results in terms of the PHQ-8 scores.

5.3.1.2 Transformer Encoder

The structure of Transformer Encoder is shown in Figure 5.11. Following [41], I use the naive Transformer Encoder structure, along with the Positional Encoding module, Multi-head Attention module, and Feed-Forward module in my work. In both the Multi-head Attention and Feed-Forward modules, data streams are designed as a shortcut structure with additive and normalization operations. An entire single Transformer Encoder layer architecture is repeated by N times to form a complete Transformer Encoder. Before being fed to the Transformer Encoder, input streams are processed by the Positional Encoding module to alter the positional information. Before being fed into the Multi-head Attention module, an input stream will be independently mapped to three sub-streams represented as Q, K, and V, respectively. Then, the Multi-head Attention module will perform global self-attention from Q, K, and V. If the head number of the Multi-head Attention module is greater than one, the Multi-head Attention module will perform the self-attention in different temporal scales. The Feed-Forward module is a simple feed-forward structure composed of two fully connected layers.

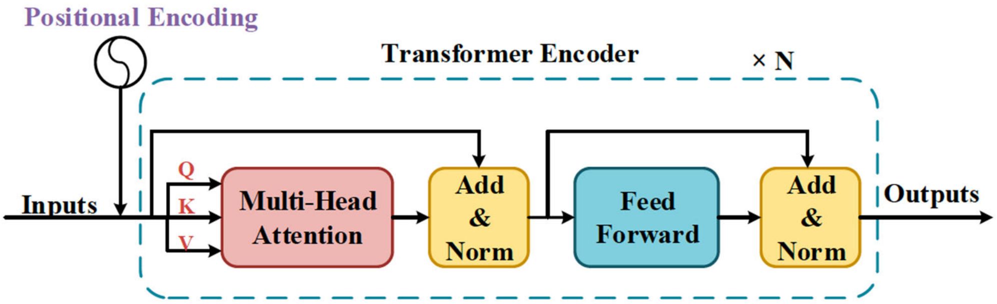


Figure 5.11: The structure of the Transformer Encoder employed to extract temporal information of sequences.#After the data processed, the data are fed to the Transformer Encoder to extract temporal information. A single Transformer Encoder layer is composed by a Multi-Head Attention module and a Feed-Forward module with an external Positional Encoding module.

The Positional Encoding module is used to add positional information to the original input. The Positional Encoding model is important because the Transformer Encoder has a pure attention structure without any positional information. The positional encoding method as [4], whose formula is shown in Equation (5.7):

$$\begin{aligned} PE(\text{ pos },2i) &= \sin\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \\ PE(\text{ pos },2i+1) &= \cos\left(\frac{\text{pos}}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \end{aligned} \tag{5.7}$$

where *pos* denotes the position; *i*, the indices of elements in every single feature; and d_{model}, the dimension of input features.

The Self-Attention module is designed to map a query (**Q**) and a set of key (**K**) -value (**V**) pairs to an attention value (Z). The **Q**, **K**, **V** are represented as Equation 5.8:

$$\begin{aligned} \mathbf{Q} &= \mathbf{W}_Q\mathbf{X} \ \in R^{F\times 1} \\ \mathbf{K} &= \mathbf{W}_K\mathbf{X} \ \in R^{E\times 1} \\ \mathbf{V} &= \mathbf{W}_V\mathbf{X} \ \in R^{E\times 1} \end{aligned} \tag{5.8}$$

where **X** represents the origin inputs, while **Q**, **K**, and **V** denote the query vector, key vector and value vector, respectively. Suppose the dimensions of **Q**, **K**, **V** are *F, E, E*, respectively. $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$ are linear transform metrices for **Q**, **K**, **V**, respectively, which are learned to find best **Q**, **K**, **V** during the training process.

In self-attention, we first calculate the similarity between **Q** and **K** as Equation 5.9:

$$\mathbf{A} = \frac{\mathbf{QK^T}}{\sqrt{E}} \in R^{F\times E} \tag{5.9}$$

where A is a similarity matrix or score matrix with a dimension of $F \times E$. Its element a_{ji} can be represented as

$$a_{ji} = q_j k_i/\sqrt{E} \tag{5.10}$$

where q_j and k_i are j-th element of Q and i-th element of K, respectively. We use softmax to normalize a_{ji} as

$$a'_{ji} = softmax(a_{ji}) = exp(a_{ji})/\sum_{j=1}^{F} exp(a_{ji}) \tag{5.11}$$

The attention value (z_j) for q_j can be represented as:

$$z_j = \sum_{i=1}^{E} a'_{ji}\, v_i \tag{5.12}$$

where v_i is the i-th element of **V**. The Z∈ $R^{F\times 1}$ can be represented as:

$$Z = A'V \tag{5.13}$$

whereA' is the normalized similarity matrix, whose element is a'_{ji}

The final feedforward module is made up of two fully connected layers whose hidden units can be specified as hyperparameters.

5.3.1.3 Multi-Task Learning

To achieve the purpose of multi-task learning, after the features are processed by the *Embedding FC Block*, the hidden embedding for each type of feature will be fed to two *FC Block*s to separately perform two tasks, i.e., PHQ-8 regression and 5-class classification. The *FC Block*s comprise a ReLU activation, a dropout layer, and a linear layer. Since we can only achieve the original PHQ-8 regression task using the AVEC 2019 DDS Challenge dataset, I generate 5-class classification labels from the original PHQ-8 score labels, as detailed in the *Data Processing* Section.

My multi-task loss function in the training stage can be formulated as:

$$\text{Loss} \ = a * L_{re} + b * L_{cl} \tag{5.14}$$

where L_{re} and L_{cl} are loss functions for PHQ-8 regression and 5-class classification, respectively. a and b in Equation (5.3) are designed to leverage the coefficient between these two tasks and can be set as hyperparameters. Specifically, the loss function for PHQ-8 regression can be formulated as follows:

$$L_{re} = 1 - \frac{2S_{\hat{y}y}}{S_{\hat{y}}^2 + S_{y}^2 + (\bar{\hat{y}} - \bar{y})^2} \tag{5.15}$$

where $\hat{y}$ and y denote the predicted depression levels and true labels, respectively. I employ the commonly used cross-entropy loss as the loss function of the 5-class classification task, which is shown as follows:

$$L_{cl} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} 1_{[c=y_i]} \log p_{i,c} \tag{5.16}$$

where *C* denotes the number of classification classes; *N*, the number of samples; $1_{[c=y_i]}$, a binary indicator; and $p_{i,c}$, the predicted probability that sample *i* belongs to class *c*.

5.3.1.4 Adaptive late fusion

To fuse results from different modalities and adjust weights for each type of feature adaptively, I employ the proposed late-fusion strategies called Adaptive Late-Fusion to fuse results obtained from every single feature.

The general late-fusion strategy that is widely used takes the average from results obtained from each feature of different modalities, which can be formulaically expressed as follows:

$$\text{FinalPredictions\_Averaged} = \frac{\sum_{m=0}^{M} \text{Predictions}_{m}}{\text{count}\,(m)} \tag{5.17}$$

where *M* denotes the number of selected features, and *Predictions*m denotes the predictions from feature *m*. In this study, the general late-fusion strategy is known as Averaged Late-Fusion. The Adaptive Late-Fusion method proposed in my work aims to increase the weights of features with high performance while decreasing the weights of features with low performance. Specifically, I calculate the weights for each feature and take the weighted average from all modalities. Weights are calculated according to the CCC from each type of feature, and thus, the feature types with higher CCC will have larger weights in the proposed Adaptive Late-Fusion. The formulaic expression of my proposed Adaptive Late-Fusion is shown as follows:

$$\begin{aligned} CCC_{Sum} &= \sum_{m=0}^{M} CCC_m \\ \text{FinalPredictions\_Weighted} &= \sum_{m=0}^{M} \frac{(\text{ Predictions}_{m} * CCC_m)}{CCC_{Sum}} \end{aligned} \tag{5.18}$$

where M denotes the number of selected features; $Predictions_{\mathrm{m}}$, the predictions from feature m; $\mathrm{CCC}_{\mathrm{sum}}$, the sum of CCCs for all features; and $\mathrm{CCC}_{\mathrm{m}}$, the CCC score of feature m.

I implement the Adaptive Late-fusion Strategy in two ways. In one way, I select all modalities and all types of features and fuse the results from them, which means that the results obtained will account for every modality. In the other way, I only fuse the top M features ranked by the CCC [52] metric, which means that I will drop features with poor performance.

5.3.2 Experiments and Results

5.3.2.1 The AVEC 2019 DDS Challenge Dataset

The DDS dataset was obtained from AVEC 2019 [26], where the level of depression (PHQ-8 questionnaire [53]) was assessed from audio-visual recordings of patients' clinical interviews conducted by a virtual agent driven by a human as a Wizard-of-Oz (WoZ) [54]. The recording audio has been transcribed by Google Cloud's speech recognition service and annotated for a variety of verbal and nonverbal features. Each interview in the AVEC 2019 DDS dataset comprises interview IDs, PHQ-8 binary labels, PHQ-8 scores, and the participant's gender. The dataset contains baseline features extracted from audiovisual recordings by common frameworks based on open-source toolboxes. It spans three expressions levels: functional low-level descriptors (hand-crafted), bag-of-words, and unsupervised deep representations. The audio features are provided by openSMILE [55], and the video features are provided by openFace [56].

For every sample in the AVEC 2019 DDS dataset, the PHQ-8 scores range ∈ [0,24]. Following [42], I define the cut-points at [0,5,10,15,20] for minimal depression, mild depression, moderate depression, moderately severe depression, and severe depression, respectively. The dataset includes MFCC, Bow_MFCC, eGeMAPS, Bow_eGeMAPS, DS_DNet, and DS_VGG for audio and FAUs, BoVW, ResNet, and VGG for video, where Bow indicates the bag-of-word method; DS, the deep spectrogram; and DNet and VGG, the data processed by pretrained DenseNet and VGG_Net, respectively. In this dataset, every modality feature has the shape of R^{t*d}, where t denotes the length of the sequence, and d represents the dimension of the modality.

5.3.2.2 Data processing

Because the data sequences are too long to fit in memory, we must shorten the dataset's original data. To shorten the sequences, I sample *N* frames from the original features for every modality feature. Specifically, I evenly split the sequence into *s* segments; for each segment, I randomly sample *L=N/s* successive frames. Finally, I concatenate the *s* segments obtained from each segment. Consequently, I can obtain *N* frames from each kind of feature in this manner. For different types of features in the AVEC 2019 DDS dataset, I select different *N* and *s*, which can be treated as hyperparameters as the dimensions of different features are different.

I generate the MFCC_functional, eGeMAPS_functional, and AUpose_functional from MFCC, eGeMAPS, and AUpose, respectively, using the approach provided by the AVEC 2019 DDS to enhance the modality and avoid the side effect of processing extremely long-term sequences. Specifically, the functional features have the same lengths as *1768* and the mean value and standard deviation of the original data.

To achieve the goal of multi-task learning, I obtain classification labels from the original PHQ-8 scores, as illustrated by [42]. The corresponding relationships between the original PHQ-8 scores and 5-class classification labels and the label distributions are presented in Table 5.12.

Table 5.12: Distribution of Training and development splits with the relationships between 5-class classification labels and PHQ-8 regression labels.

Task		Train	Dev
Regression Task		163	56
Classification Task	minimal [0-4]	77	26
	mild [5-9]	36	15
	moderate [10-14]	26	8
	moderately severe [15-19]	17	6
	severe [20-24]	7	1

5.3.3 Evaluation Functions

5.3.3.1 Data processing

I use well-known standard evaluation metrics for depression detection to evaluate regression/classification results. I use the Concordance Correlation Coefficient (CCC) [52] as a measure of PHQ-8 estimated scores (regression task), which is the common metric in dimensional depression detection to measure the agreement between true PHQ-8 scores (y) and predicted PHQ-8 scores ($\hat{y}$). The CCC is formulated as follows:

$$CCC = \frac{2S_{\hat{y}y}}{S_{\hat{y}}^2 + S_y^2 + (\bar{\hat{y}} - \bar{y})^2} \tag{5.19}$$

where $S_{\hat{y}}$ and S_y denote the variances of $\hat{y}$ and y, whereas $S_{\hat{y}y}$ denote the corresponding covariance value. The CCC is based on Lin's calculation [52]. The range of the CCC is from -1 to 1, where -1 represents perfect disagreement and 1 represents perfect agreement.

As another measure for the regression task, I also use the root mean square error (RMSE), which is defined as Equation (5.9), where $\hat{y}$ and y denote the predicted and true depression levels, respectively, and N represents the number of samples.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i^N (y_i - \hat{y}_i)}{N}} \tag{5.20}$$

5.3.4 Experimental Setup

To demonstrate the effectiveness of the proposed method, I apply it, along with the original baseline GRU model [26], to obtain a direct comparison between them. The AVEC 2019 DDS dataset is split into training, development, and test sets. I utilized only the training and development sets for a fair comparison with the state-of-the-art method. The experiments were conducted on 219 subjects: 163 subjects for training and 56 subjects for development. The Adam optimization algorithm [57] was employed to learn the parameters in the networks. The learning rate was set to 1e-5. The batch size was set to 48 for low- and middle-level features and 24 for high-level features. I trained the model for 500 epochs for low&middle-level features and 200 epochs for high-level features. During training, I set *a, b* to 1.0, 0.0 for single-task and 0.9, 0.1 for multi-task in the loss function of Equation (5.3). For the Transformer Encoder block, I set the head number of Multi-head Attention to 1, the hidden dimension of the Feed-Forward layer to 2048, and the number of the encoder layer to 6 following the original Transformer structure [41]. The model is implemented with the framework of PyTorch [58], whereas the experiments are conducted on double Nvidia RTX 3090 GPU cards.

5.3.5 Results

In this section, I will describe the results of the experiments in more detail. The proposed networks have several hyperparameters to be optimized. The length of inputs for the Transformer Encoder *N* and the number of selected modalities for fusion *M* is the most important architectural decisions. In this section, I first discuss the effect of the selection of *N* and *M*. Then, I describe the effectiveness of the Transformer Encoder, multi-task learning, and multi-modal learning. Finally, I compare the proposed models' results with those of some state-of-the-art methods.

5.3.5.1 Effect of Frames Length

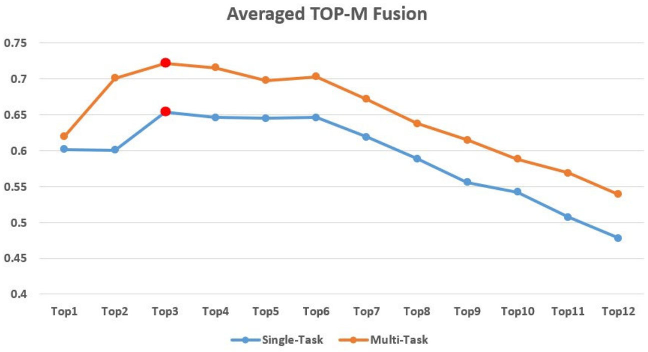
The length of inputs (i.e., the number of sequence frames) affects the accuracy of the depression detection of the networks and should thus be selected carefully. To study how the performance of the proposed method changes as I modify the input frames length, I fix the task as regression and modalities as the fusion of top three features (*M=3*) and compare the CCC score and RMSE at different selections of frames. As feature dimensions significantly differ, I select the same frames for low- and middle-level features, whereas I select different frames for high-level features. Although the use of the transformer model can capture long-term information, it consumes a lot of memory. The pair of *2048/720* frames for low&middle-level features and high-level features is the limitation of our hardware. As presented in Table 5.13, an increase in input frames improves the results. I select *N = 2048* for low- and middle-level features and *N = 720* for high-level features in this work. Here *{N}* is the number of frames.

Table 5.13: The CCC and RMSE results for different frames of input on fixed 3-top modalities fusion regression single task. 2048/720 means that I select N=2048 for low-level and middle-level features and N=720 for high-level features.

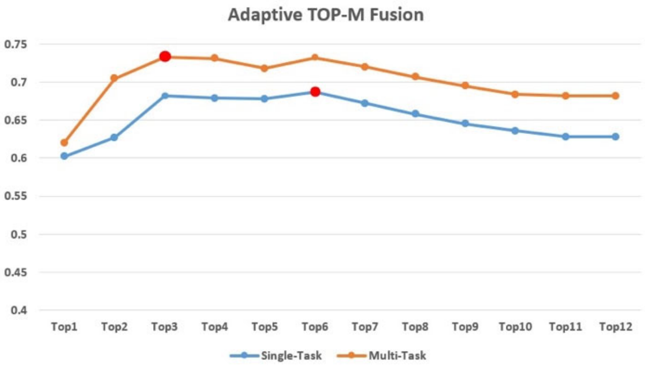
Frames(N) (low&middle-level features/high-level features)	CCC	RMSE
2048/720	**0.654**	**4.602**
1536/540	0.634	4.526
1024/360	0.560	5.102

5.3.5.2 Effect of the Selection of Features

To study the effect of feature selection in the late-fusion, Figure 5.12 presents the results of different *M* selection in terms of the CCC scores with different multi-task combinations. The most appropriate number of selected top modalities for different tasks varies. The results indicate that the CCC scores increase with an increase in the number of top modalities selected and reach a maximum at $M = 3$ for both single-task and multi-task in Averaged Late-Fusion. $M = 3$ and $M = 6$ are the best choices for multi-task and single-task in Adaptive Late-Fusion, respectively. Therefore, I select the corresponding best *M* for different tasks in different fusion strategies.



(a) Averaged Top-M Fusion



(b) Adaptive Top-M Fusion

Figure 5.12: The CCC scores for different number of top M modalities fusion. Each color represents different tasks. The points with best results are marked as red.

5.3.5.3 GRU vs. Transformer Encoder

To investigate the effect of transformer-based networks on the CCC scores and RMSE values, I use single features as inputs of the networks, and the task is set to PHQ-8 regression. The results are presented in Table 5.14.

Table 5.14: Comparison between The GRU baseline model [26] and The Transformer-based model for different features from audio and video modality. For every kind of features from each modality, I use two metrics including CCC and RMSE.

Feature		CCC		RMSE	
		GRU [26]	Proposed method	GRU [26]	Proposed method
Low-level	MFCC	0.198	**0.289**	7.28	**5.70**
	MFCC_functional	-	**0.386**	-	**7.78**
	eGeMAPs	**0.076**	0.0002	**7.78**	8.69
	eGeMAPs _ functional	-	**0.138**	-	**8.10**
	AUposes	0.115	**0.602**	7.02	**5.64**
	AUposes_functional	-	**0.277**	-	**6.23**
Middle-level	BoW-MFCC	**0.102**	0.060	**6.32**	8.58
	BoW-eGeMAPs	**0.272**	0.169	**6.43**	8.56
	BoW-AUposes	0.107	**0.210**	**5.99**	9.045
High-level	DeepSpectrogram_DNet	0.165	**0.204**	**8.09**	8.67
	DeepSpectrogram_VGG	**0.305**	0.141	**8.00**	8.72
	Facial_ResNet	0.269	**0.373**	7.72	**7.56**

As presented in Table 5.14, the transformer model outperforms the GRU baseline model [26] in terms of the CCC metric for low- and high-level features. The CCC score of the AUpose feature is higher than those of other types of features. The transformer-based network achieves higher accuracy for low&high-level features, so we can conclude that the transformer model outperforms the GRU [26] in terms of processing low- and high-level features. However, for middle-level features, we can deduce that the transformer-based network does not outperform the GRU model.

5.3.5.4 Single-Task vs. Multi-Task

To illustrate the effectiveness of multi-modal learning, the proposed method has been tested on all features available in the AVEC 2019 DDS dataset. As presented in Table 5.15, applying multi-modal late-fusion outperforms any unimodal learning in any tasks, except for the Averaged All Late-Fusion. The Averaged All Late-Fusion has poor performance as it does not take the importance of different modalities into account.

Table 5.15: Comparison between single-task and multi-task with metrics of CCC, RMSE. Single-Task includes the PHQ-8 regression task while Multi-Task includes the PHQ-8 regression task and the 5-class classification task.

Tasks	CCC	RMSE
Our: Single-Task	0.679	4.150
Our: Multi-Task	**0.733**	**3.783**

5.3.5.5 Single Modality vs. Averaged Multi-modal Fusion vs. Adaptive Multi-modal Fusion

To illustrate the effectiveness of multi-modal learning, the proposed method has been tested on all features available in the AVEC 2019 DDS dataset. As presented in Table 5.16, applying multi-modal late-fusion outperforms any unimodal learning in any tasks, except for the Averaged All Late-Fusion. The Averaged All Late-Fusion has poor performance as it does not take the importance of different modalities into account.

Table 5.16: Comparison between single modality and multiple modalities fusion and Comparison between averaged multi-modal fusion and adaptive multi-modal fusion. I employ CCC and RMSE as metrics. For every kind of fusion strategy, I perform two ways of late fusion including *all fusion* and *top M fusion*.

	CCC		RMSE	
	Single-Task	Multi-Task	Single-Task	Multi-Task
MFCC	0.289	0.471	5.700	5.158
MFCC_functional	0.386	0.460	7.780	7.269
eGeMAPs	0.0002	0.000	8.688	8.688
eGeMAPs_functional	0.138	0.107	8.102	8.345
AUposes	0.602	0.620	5.643	5.355
AUposes_functional	0.277	0.390	6.227	7.114
BoW_MFCC	0.060	0.063	8.584	11.575
BoW_eGeMAPs	0.169	0.181	8.560	8.555
BoW_AUposes	0.210	0.184	9.045	10.760
DeepSpectrogram_Dnet	0.204	0.171	8.672	8.662
DeepSpectrogram_VGG	0.141	0.170	8.721	8.145
Facial_ResNet	0.373	0.360	7.561	6.900
Averaged all fusion	0.478	0.539	4.591	4.334
Averaged best top M fusion	0.654	0.722	4.602	3.852
Adaptive all fusion	0.628	0.682	4.046	**3.782**
Adaptive best top M fusion	**0.687**	**0.733**	**3.829**	3.783

The result indicates that the fusion of the best top M modalities improves the estimation of depression levels in terms of the CCC metric better than other fusions as modalities with poor performance are excluded to avoid a negative impact on the accuracy of depression detection. We can infer that the Adaptive Late-Fusion strategy can perform better than the Average Late-Fusion in estimating the levels of depression.

To investigate the different weights between different features, I counted the best three features with their corresponding weights in Adaptive Late-Fusion because *M = 3* is the best choice for multi-task learning and nearly the best choice for single-task learning. As presented in Table 5.17, although the selections of modalities are different for different tasks, the main influencing features are AUposes and MFCC_Functional. The results indicate that low-level features are more important than deep- and middle-level features for estimating the levels of depression.

Table 5.17: The selection of modalities in TOP-3 Adaptive Late-Fusion and their corresponding weights.

Tasks	Best 3 features and corresponding weights			
Single-Task	Modality	Video	Audio	
	Features	AUposes	MFCC	MFCC_Functional
	Weights	0.40	0.30	0.30
Multi-Task	Modality	Video		Audio
	Features	AUPoses	ResNet	MFCC_Functional
	Weights	0.44	0.27	0.29

5.3.5.6 Comparison with state-of-the-art methods

In Table 5.18, my approaches are compared with other state-of-the-art methods and the baseline. The baseline model [26] uses a GRU to extract temporal information and then takes the average of the results from every unimodality. The hierarchical Bi-LSTM [39] hierarchically employs a Bi-LSTM to obtain temporal sequence information. Multi-scale temporal dilated CNN [40] employs dilated CNNs with different scales to process temporal information, followed by average pooling, and max pooling to fuse temporal features. It should be noted that multi-scale temporal dilated CNN [40] uses features from texts extracted from pretrained models. Bert-CNN & Gated-CNN [49] use the Gated-CNN to extract features from each audio-visual modality sequence and the Bert-CNN to obtain features from texts before fusing the features to predict the final depression levels. The results indicate that the baseline has superior performance over other DL methods.

Table 5.18: Comparation of the proposed method and the state-of-the-art with CCC metrics and modalities used.

Methods	CCC	Modalities Used
Baseline [26]	0.336	Audio/Video
Hierarchical BiLSTM [39]	0.402	Audio/Video/Text
Multi-scale Temporal Dilated CNN [40]	0.466	Audio/Video/Text
Bert-CNN & Gated-CNN [49]	0.696	Audio/Video/Text
Ours Best	**0.733**	Audio/Video

5.3.6 Discussion

Compared with Average Late-Fusion, Figure5.12 shows that using Adaptive Late-Fusion not only achieves good results but also increases detection robustness, implying that the inclusion of low-performance features has a slightly negative impact on the detection results.

The comparison of the predicted results with the ground truth is presented in Figure 5.13, and samples with different classification labels are colored differently. As shown in Figure 5.13, the predicted results of participants with high scores tend to be on the lower side. The reason is due to the imbalance of the training samples. Figure 5.14 shows the distribution of the training set of the AVEC 2019 DDS. The training set distribution is unbalanced, more samples have participants with low PHQ-8 scores, whereas few samples have participants with high scores. As a result, the model predicts a slightly lower PHQ-8 score than the true label for participants with high scores. The prediction accuracy can be improved by increasing the number of participants with high scores.

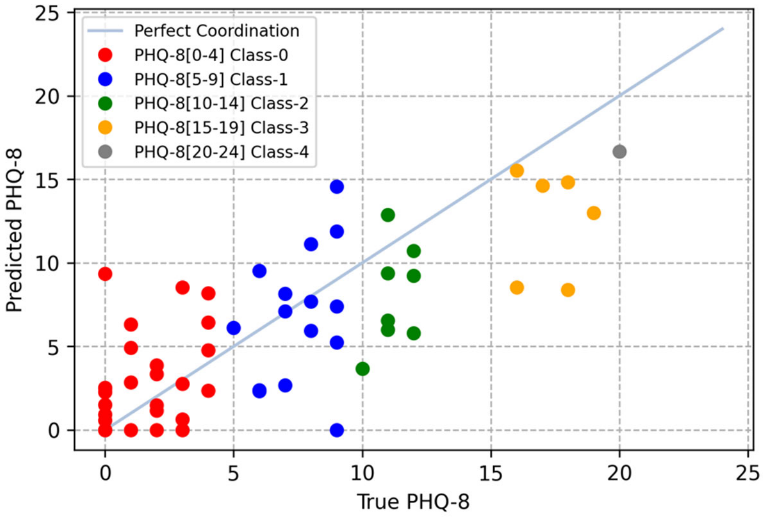


Figure 5.13: Correlation graph between the predicted and true PHQ-8 scores. Each color represents different classes.

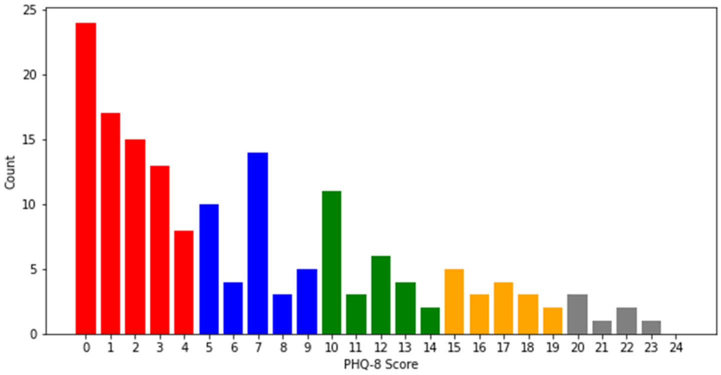


Figure 5.14: The distribution of the training set of the AVEC 2019 DDS Challenge dataset.

5.4 Chapter Summary

In this Chapter, I introduced a novel multimodal behavioral dataset of depressive severity. I have also presented the baseline networks and their results for audio and visual features. These results indicated that low level features performed better in audio-based depression detection while deep learning features performed better in visual based depression detection. The prediction results in emotional speech scenario indicated behavioral features in positive-emotional speech have more potential in depressive severity identification.

I also presented a multi-modal adaptive fusion transformer network for depression detection using multi-task representation learning with facial and acoustic features, which achieves the best results on the development set of the AVEC 2019 DDS dataset. The experimental results indicated that the use of the transformer model for depression detection can improve the final prediction performance. The ablation study demonstrated that multi-task representation learning, with tasks such as PHQ-8 regression and 5-class classification, outperforms single-task representation learning for depression detection. However, the results indicated that the combination of the regression task and the binary gender classification task cannot outperform the combination of the regression task and the 5-class classification task. The experimental results indicated that Adaptive Late-Fusion contributes more significantly than Averaged Late-Fusion to depression detection performance while also increasing robustness. By fusing the selected modalities, the proposed approach achieved a CCC score of 0.733 on the AVEC 2019 DDS dataset, outperforming the alternative methods investigated in this work.

Bibliography

1. Chen, L., Wang, L., Qiu, X. H., Yang, X. X., Qiao, Z. X., Yang, Y. J., & Liang, Y.: Depression among Chinese university students: prevalence and socio-demographic correlates. PloS one, 8(3), e58379 (2013).
2. Lei, X. Y., Xiao, L. M., Liu, Y. N., & Li, Y. M.:Prevalence of depression among Chinese University students: a meta-analysis. PLoS One, 11(4), e0153454(2016).
3. Setterfield, M., Walsh, M., Frey, A. L., & McCabe, C.: Increased social anhedonia and reduced helping behaviour in young people with high depressive symptomatology. Journal of Affective Disorders, 205, 372-377(2016).
4. Brinkmann, K., & Franzen, J.:Blunted cardiovascular reactivity during social reward anticipation in subclinical depression. International Journal of Psychophysiology, 119, 119-126 (2017).
5. American Psychiatric Association.: Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub (2013).
6. Hysenbegasi, A., Hass, S. L., & Rowland, C. R.: The impact of depression on the academic productivity of university students. Journal of Mental Health Policy and Economics, 8(3), 145(2005).
7. Hu, T. W.: The economic burden of depression and reimbursement policy in the Asia Pacific region. Australasian Psychiatry, 12(sup1), s11-s15(2004).
8. Sobocki, P., Lekander, I., Borgström, F., Ström, O., & Runeson, B.: The economic burden of depression in Sweden from 1997 to 2005. European Psychiatry, 22(3), 146-152(2007).
9. Aalto-Setälä, T., Marttunen, M., Tuulio-Henriksson, A., Poikolainen, K., & Lö-nnqvist, J.: Depressive symptoms in adolescence as predictors of early adulthood depressive disorders and maladjustment. American Journal of Psychiatry, 159(7), 1235-1237(2002).
10. Liu, X. C., Ma, D. D., Kurita, H., & Tang, M. Q.: Self-reported depressive symptoms among Chinese adolescents. Social psychiatry and psychiatric epidemiology, 34(1), 44-47 (1999).

11. Jan, A., Meng, H., Gaus, Y. F. B. A., & Zhang, F.: Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. IEEE Transactions on Cognitive & Developmental Systems, PP(99), 1-1 (2017).
12. Jan, A., Meng, H., Gaus, Y. F. A., Zhang, F., & Turabzadeh, S.: Automatic depression scale prediction using facial expression dynamics and regression. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (pp. 73-80). ACM (2014, November).
13. Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., & Sahli, H.: Multimodal measurement of depression using deep learning models. In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (pp. 53-59). ACM(2017, October).
14. Dibeklioğlu, H., Hammal, Z., & Cohn, J. F.: Dynamic multimodal measurement of depression severity using deep autoencoding. IEEE journal of biomedical and health informatics, 22(2), 525-536 (2018).
15. Girard, J. M., Cohn, J. F., Mahoor, M. H., Mavadati, S., & Rosenwald, D. P.: Social risk and depression: Evidence from manual and automatic facial expression analysis. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (pp. 1-8). IEEE (2013, April).
16. Wang, Z., Yuan, C. M., Huang, J., Li, Z. Z., Chen, J., Zhang, H. Y., ... & Xiao, Z. P.: Reliability and validity of the Chinese version of Beck Depression Inventory-II among depression patients. Chinese Mental Health Journal, 25(6), 476-480 (2011).
17. Beck, A. T., Steer, R. A., & Brown, G. K.: Beck depression inventory-II. San Antonio, 78(2), 490-498 (1996).
18. Wang, X. D., Wang, X. L., Ma, H.: Manual of mental health assessment scales. Chinese Mental Health Journal(supplement), 1999.
19. Radloff, & L., S.: The CES-D scale: a self-report depression scale for research in the general population. Applied Psychological Measurement, 1(3), 385-401(1977).
20. Hamilton, M.: Development of a rating scale for primary depressive illness. British Journal of Clinical Psychology, 6(4), 278-296 (1967).
21. Liu, J. Q., Huang, Y., Huang, X. Y., Xia, X. T., Niu, X. X., Chen, Y. W.: Multimodal Behavioral Dataset of Depressive Symptoms in Chinese College Students–Preliminary

Study. In: Chen YW. et. al. (eds) Innovation in Medicine and Healthcare Systems, and Multimedia. Smart Innovation, Systems and Technologies, vol 145. Springer, Singapore, pp. 79-190, (2019).

22. McIntyre, G., Göcke, R., Hyett, M., et al. An approach for automatically measuring facial activity in depressed subjects[C]//2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. IEEE, 1-8 (2009).
23. Pan, W, Wang, J, Liu, T, et al. Depression recognition based on speech analysis[J]. Chinese Science Bulletin, 63(20): 2081-2092, (2018).
24. Wang, J. Y. An Exploratory Study on Auxiliary Diagnosis of Depression Based on Speech. Doctoral Dissertation. Beijing: Chinese Academy of Science, (2017).
25. Joshi, J., Goecke, R., Parker, G., & Breakspear, M.: Can body expressions contribute to automatic depression analysis?. In 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) (pp. 1-7). IEEE (2013, April).
26. Ringeval F., Schuller B., Valstar M., Cummins N., Cowie R., Tavabi L., Schmitt M., Alisamir S., Amiriparian S., Messner E.-M., et al.: Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, pages 3–12, 2019.
27. Florian E., Felix W., Florian G., and Björn S.: Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. In Proc. 21st ACM International Conference on Multimedia (ACM MM). ACM, Barcelona, Spain, 835–838, 2013.
28. Florian E, Klaus R. S., Björn S., Johan S., Elisabeth A., Carlos B., Laurence D., Julien E., Petri L., Shrikanth S. N., and Khiet P. T. :The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing 7, 2 (April 2016), 190–202.
29. Maximilian S. and Björn S.: openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit. Journal of Machine Learning Research 18, 96 (2017), 1–5.
30. Karen S. and Andrew Z.: Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556. 14 pages, 2014.

31. Gao H., Zhuang L., Laurens van der M., and Kilian Q. W.: Densely Connected Convolutional Networks. In The IEEE Conference on Computter Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HW, 4700–4708,2017.

32. Deng, J. and Dong, W. and Socher, R. and Li, L.-J. and Li, K. and Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database, In The IEEE Conference on Computter Vision and Pattern Recognition (CVPR), 2009.

33. Tadas B., Amir Z., Yao C. L., and Louis-P. M.: OpenFace 2.0: Facial Behavior Analysis Toolkit. In Proc. 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, Xi'an, P. R. China, 59–66, 2018.

34. Dimitrios K., Panagiotis T., Mihalis A. N., Athanasios P., Guoying Z., Björn S., Irene K., and Stefanos Z.: Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. International Journal of Computer Vision 127, 6 (2019), 907–929, 2019.

35. Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014a). Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR, abs/1406.1078, 2014.

36. Chung J., Gülçehre C.,Cho K., and Bengio Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555, 2014.

37. Lawrence I-Kuei L.: A concordance correlation coef- ficient to evaluate reproducibility, Biometrics, pp. 255– 268, 1989.

38. Liu, J.Q.; Huang, Y.; Huang, X.Y.; Xia, X.T.; Niu, X.X.; Chen, Y.W. Multimodal behavioral dataset of depressive symptoms in chinese college students–preliminary study. In Innovation in Medicine and Healthcare Systems, and Multimedia; Springer, 2019; pp. 179–190.

39. Yin, S.; Liang, C.; Ding, H.; Wang, S. A multi-modal hierarchical recurrent neural network for depression detection. Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 65–71.

40. Fan, W.; He, Z.; Xing, X.; Cai, B.; Lu, W. Multi-modality depression detection via multi-scale temporal dilated cnns. Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 73–80.

41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.;

Polosukhin, I. Attention is all you need. arXiv preprint arXiv:1706.03762 2017.

42. Qureshi, S.A.; Saha, S.; Hasanuzzaman, M.; Dias, G. Multitask representation learning for multimodal estimation of depression level. IEEE Intelligent Systems 2019, 34, 45–52.
43. Wang, Y.; Wang, Z.; Li, C.; Zhang, Y.; Wang, H. A Multitask Deep Learning Approach for User Depression Detection on Sina Weibo. arXiv preprint arXiv:2008.11708 2020.
44. Keras: https://Kreas.io/
45. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han,W.;Wang, S.; Zhang, Z.;Wu, Y.; others. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 2020.
46. Delbrouck, J.B.; Tits, N.; Brousmiche, M.; Dupont, S. A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis. arXiv preprint arXiv:2006.15955 2020.
47. Anderson, I.M.; Shippen, C.; Juhasz, G.; Chase, D.; Thomas, E.; Downey, D.; Toth, Z.G.; Lloyd-Williams, K.; Elliott, R.; Deakin, J.W.State-dependent alteration in face emotion recognition in depression. The British Journal of Psychiatry 2011, 198, 302–308.
48. Joshi, J.; Goecke, R.; Alghowinem, S.; Dhall, A.; Wagner, M.; Epps, J.; Parker, G.; Breakspear, M. Multimodal assistive technologies for depression diagnosis and monitoring. Journal on Multimodal User Interfaces 2013, 7, 217–228.
49. Rodrigues Makiuchi, M.; Warnita, T.; Uto, K.; Shinoda, K. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, pp. 55–63.
50. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018.
51. Kroenke, K.; Spitzer, R.L. The PHQ-9: a new depression diagnostic and severity measure, 2002.
52. Lawrence, I.; Lin, K. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989, pp. 255–268.
53. Kroenke, K.; Strine, T.W.; Spitzer, R.L.; Williams, J.B.; Berry, J.T.; Mokdad, A.H. The

PHQ-8 as a measure of current depression in the general population. Journal of affective disorders 2009, 114, 163–173.

54. Gratch, J.; Artstein, R.; Lucas, G.M.; Stratou, G.; Scherer, S.; Nazarian, A.; Wood, R.; Boberg, J.; DeVault, D.; Marsella, S.; others. The distress analysis interview corpus of human and computer interviews. LREC, 2014, pp. 3123–3128.
55. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462. Version May 25, 2021 submitted to Sensors 14 of 15
56. Baltrušaitis, T.; Robinson, P.; Morency, L.P. Openface: an open source facial behavior analysis toolkit. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016, pp. 1–10.
57. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.
58. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; others. Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 2019.
59. Parkhi, OM., Vedlda, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference 2015, Section 3, pp. 41.1–41.12 (2015)
60. Afouras, T., Chung, JS., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
61. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Proceedings of the Asian Conference on Computer Vision (2016)
62. Yuan, J., Liberman, M.: Speaker identification on the SCOTUS corpus. J. Acoust. Soc. Am. 123(5), 3878 (2008)
63. Dlib: http://dlib.net/
64. Alghowinem S, Goecke R, Wagner M, Epps J, Breakspear M, Parker G. From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech. Proc FLAIRS-25. 2012:141–146.
65. Yang Y, Fairbairn CE, Cohn JF. Detecting depression severity from intra- and interpersonal vocal prosody. IEEE Trans on Affective Computing. 2013

66. Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum DR. The Distress Analysis Interview Corpus of human and computer interviews. InLREC 2014 May (pp. 3123-3128)

67. Bell, C. C. DSM-IV: diagnostic and statistical manual of mental disorders. JAMA 272, 828–829 (1994).

Chapter 6

Conclusion

In this research, I presented some generation, representation, and fusion methods for multi-modal deep learning and demonstrate state-of-art performance on three datasets that span the task domains of hand gesture recognition, human pose recognition, and estimation of depression level.

The achievements are summarized as follows:

- Firstly, I built a new multi-angle RGB-D dataset (MaHG-RGBD) with 15 participants performing 25 hand gestures. Not only the front-view but also the tilted view (titled angle = 45 degrees) dataset are provided, which can be used when space is limited especially in the surgery room. Based on the multimodal RGB-D dataset, I primarily focus on proposed a multimodal deep learning method to perform recognition hand gestures using color and depth information.
- Second, for the touchless interaction systems for visualization of hepatic anatomical models in surgery, I have proposed four versions. In the first version, I used HOG as features and SVM as a classifier to recognize 9 kinds of hand gestures from the depth images, the average recognition accuracy is found to be 87.5% with the speed of 8fps. Though the HOG-based machine learning method can recognize various hand gestures with reasonable accuracy, they could not achieve real-time recognition. In the second version, the system uses a Kinect sensor to acquire three kinds of hand states and track hand their movements. Based on these states and their movements, I designed a range of hand gestures, and finally, four kinds of operations are available using touchless gestures to visualize 3D hepatic anatomic models in real-time. Though this version is a prototype, the preliminary result is encouraged. For the third version, I develop a deep learning technique for recognition of various hand gestures to increase the degree of freedom of operations and achieve more flexible touchless visualization. For the fourth version, I

proposed a multimodal deep learning method to perform recognition using color and depth images. The multimodal system achieves better real-time robust recognition than conventional methods.

- Third, in my previous study, I demonstrated that depth images provide higher recognition than the color image. Though the depth image is more useful and accurate for posture recognition than the color image, the depth cameras are not as widely used and affordable as color cameras. I proposed an RGB posture-recognition network based on a two-stage CNN architecture. To improve the recognition performance from color images, I generated an estimated depth posture image by a hybrid loss function incorporated in the generation module. The loss function captures the high-level features and recovers the sharp depth discontinuities. The proposed method was evaluated on our novel dataset of color-depth pose images and the public OUHANDS hand gesture dataset. The hybrid loss effectively and accurately generated depth posture images and the estimated depth image improved the accuracy of posture recognition.
- Fourth, I built 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 102 subjects with different depression levels, which are recorded by two video cameras and five microphones. I proposed a deep learning model for depressive symptoms detection, which extracts and fuses the dynamic facial features associated with different emotion voice stimuli. The effectiveness of the proposed method is validated on the original multimodal behavioral dataset. The results demonstrated that dynamic facial expression features can potentially reveal depressive symptoms. The detection accuracy of three different emotion states (three different emotion voice stimuli) is about 71.4%. Compared with the single emotion feature, the fused multiple emotion features can significantly improve the detection accuracy. The mean accuracy was improved to 76.1%.
- Fifth, I presented a multi-modal adaptive fusion transformer network for depression detection using multi-task representation learning with facial and acoustic features, which achieves the best results on the development set of the AVEC 2019 DDS dataset. The experimental results indicated that the use of the transformer model for depression detection can improve the final prediction performance. The ablation study demonstrated

that multi-task representation learning, with tasks such as PHQ-8 regression and 5-class classification, outperforms single-task representation learning for depression detection. However, the results indicated that the combination of the regression task and the binary gender classification task cannot outperform the combination of the regression task and the 5-class classification task. The experimental results indicated that Adaptive Late-Fusion contributes more significantly than Averaged Late-Fusion to depression detection performance while also increasing robustness. By fusing the selected modalities, the proposed approach achieved a CCC score of 0.733 on the AVEC 2019 DDS dataset, outperforming the alternative methods investigated in this work.

Publication List

Journal Papers

1. **Jia-Qing Liu**, Ryoma Fujii, Tomoko Tateyama, Yutaro Iwamoto, and Yenwei Chen, "Kinect-Based Gesture Recognition for Touchless Visualization of Medical Images," *International Journal of Computer and Electrical Engineering,* vol. 9, no. 2, pp. 421-429, 2017.

2. **Jia-Qing Liu**, Kotaro Furusawa, Tomoko Tateyama, Yutaro Iwamoto, and Yen-wei Chen, "An Improved Kinect-Based Real-Time Gesture Recognition Using Deep Convolutional Neural Networks for Touchless Visualization of Hepatic Anatomical Mode," *Journal of Image and Graphics,* Vol. 7, no. 2, pp. 45-49, 2019.

3. **Jia-Qing Liu,** Tomoko Tateyama, Yutaro Iwamoto, and Yen-Wei Chen, "A Preliminary Study of Kinect-Based Real-Time Hand Gesture Interaction Systems for Touchless Visualizations of Hepatic Structures in Surgery," *Medical Imaging and Information Sciences*, Vol. 36, no. 3, pp. 128-135, 2019.（日本医用画像情報学会金森賞）

4. Hao Sun, **Jia-Qing Liu**, Shu-Rong Chai, Zhaolin Qiu, Lan-Fen Lin, Xinyin Huang and Yen-Wei Chen, Multi-modal Adaptive Fusion Transformer Network for the Estimation of Depression Level, Sensors, Vol.21, 4764, 2021. (https://doi.org/10.3390/s21144764). (co-first authors)

5. **劉 家慶**, 黄 慧敏, 健山 智子, 岩本 祐太郎, 林 蘭芬, 陳 延偉, タッチベースインタラクティブ COVID-19 診断支援可視化システム, 電子情報通信学会システム開発論文特集, 条件付採録 2020.0818.

Book

1. **Jia-Qing Liu**, Yue Huang, Shu-Rong Chai, Hao Sun, Xin-Yin Huang, Lanfen Lin, Yen-Wei Chen, (Chapter) Computer-aided Detection of Depressive Severity Using Multimodal Behavioral Data, *Handbook of Artificial Intelligence in Healthcare*, Springer, 2021 (In press)

2. Zhaolin Qiu, Lanfen Lin, Hao Sun, **Jia-Qing Liu**, Yen-Wei Chen, Artificial Intelligence in Remote Photoplethysmography: Remote Heart Rate Estimation from Video Images, Handbook of Artificial Intelligence in Healthcare, Springer, 2021 (In press)

Referred International Conference Papers

1. **Jia-Qing Liu**, Ryoma Fujii, Tomoko Tateyama, Yutaro Iwamoto, Yen-Wei Chen: "Kinect-Based Gesture Recognition for Touchless Visualization of Medical Images," *2017 4th International Conference on Mechanical, Electronics and Computer Engineering (CMECE 2017)*, Phnom Penh, Cambodia, Sep. 14-16, 2017. **Best Student Presentation Award.**

2. **Jia-Qing Liu,** Tomoko Tateyama, Yutaro Iwamoto and Yen-Wei Chen, "Kinect-Based Real-Time Gesture Recognition Using Deep Convolutional Neural Networks for Touchless Visualization of Hepatic Anatomical Models in Surgery". *Intelligent Interactive Multimedia Systems and Services. KES-IIMSS-18 2018. Smart Innovation, Systems and Technologies,* vol 98. Springer, Cham. Gold Coast, Australia, June 20-22, 2018.

3. **Jia-Qing Liu,** Kotaro Furusawa, Tomoko Tateyama, Yutaro Iwamoto, and Yen-wei Chen, "An Improved Kinect-Based Real-Time Gesture Recognition Using Deep Convolutional Neural Networks for Touchless Visualization of Hepatic Anatomical Mode," *Proc. of International Conference on Digtal Medicine and Image Processing (DMIP2018)*, pp.56-60, Okinawa, Japan, Nov.12-14, 2018. **Best Student Presentation Award.**

4. **Jia-Qing Liu**, Kotaro Furusawa, Seiju Tsujinaga, Tomoko Tateyama, Yutaro Iwamoto, Yen-Wei Chen, "MaHG-RGBD: A Multi-angle View Hand Gesture RGB-D Dataset for Deep Learning Based Gesture Recognition and Baseline Evaluations," *Proc. of IEEE 37th International Conference on Consumer Electronics (IEEE ICCE2019)*, Las Vegas, USA, Jan. 11-13, 2019.

5. **Jia-Qing Liu,** Yue Huang, Xin-Yin Huang, Xiao-Tong Xia, Xi-Xi Niu and Yen-Wei Chen, "Multimodal Behavioral Dataset of Depressive Symptoms in Chinese College

Students–Preliminary Study," In: Chen YW., Zimmermann A., Howlett R., Jain L. (eds) *Innovation in Medicine and Healthcare Systems, and Multimedia. Smart Innovation, Systems and Technologies*, vol 145. Springer, Singapore, pp.179-190, 2019 Proc. of InMed2019, Malta, June 17-19, 2019.

6. **Jia-Qing Liu,** Kotaro Furusawa, Tomoko Tateyama, Yutaro Iwamoto, *Yen-Wei Chen, "An Improved Hand Gesture Recognition with Two-Stage Convolutional Neural Networks Using a Hand Color Image and Its Pseudo-Depth Image," *Proc. of 2019 IEEE International Conference on Image Processing (IEEE ICIP 2019)*, Taibei, Taiwan, pp.375-379, Sep. 22-25, 2019.

7. **Jia-Qing Liu,** Yue Huang, Xin-Yin Huang, Xiao-Tong Xia, Xi-Xi Niu, Lanfen Lin, and Yen-Wei Chen, "Dynamic Facial Features in Positive-Emotional Speech for Identification of Depressive Tendencies" in Y.-W. Chen et al. (eds.), *Innovation in Medicine and Healthcare, Smart Innovation, Systems and Technologies 192 (Proc. of InMed2020)*, pp.127-134 (2020).

8. Seiju Tsujinaga, Nobuo Yamaguchi, **Jia-Qing Liu**, Tomoko Tateyama, Yutaro Iwamoto and Yen-Wei Chen, "Interactive Virtual Campus Tour System Using Skeleton Information from Kinect," *Proc. of 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE 2018),* Nara, Japan, Oct.9-12, 2018.

9. Kotaro Furusawa, **Jia-Qing Liu**, Seiju Tsujinaga, Tomoko Tateyama, Yutaro Iwamoto, Yen-Wei Chen, "Robust Hand Gesture Recognition Using Multimodal Deep Learning for Touchless Visualization of 3D Medical Images," In: Liu Y., Wang L., Zhao L., Yu Z. (eds) Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery. ICNC-FSKD 2019. Advances in Intelligent Systems and Computing, vol 1074. Springer, Cham, pp.593-600, 2020 (Kumin, China, July 20-22, 2019).

10. Yuan Zhuang, Lanfen Lin, Ruofeng Tong, **Jia-Qing Liu**, Yutaro Iwamot, Yen-Wei Chen, "G-GCSN: Global Graph Convolution Shrinkage Network for Emotion Perception from Gait," In: Sato I., Han B. (eds) Computer Vision – ACCV 2020 Workshops. ACCV 2020. Lecture Notes in Computer Science, Springer, vol 12628, pp.46-57, 2021. https://doi.org/10.1007/978-3-030-69756-3_4.

Symposiums in Japan

1. **Jiaqing Liu**, Ryoma Fujii, Tomoko Tateyama, Yutaro Iwamoto and Yen-Wei Chen: “Kinect-Based Gesture Recognition for Touchless Visualization of Medical Images,” 第 36 回日本医用画像工学会大会，OP2-3，p.32，岐阜，2017.7.27-29.

2. **Jiaqing Liu**, Kotaro Furusawa, Seiju Tsujinaga, Tomoko Tateyama, Yutaro Iwamoto, Yen-Wei Chen, “Kinect RGB-D Hand Gesture Image Database for Deep Learning Based Gesture Recognition,” 電子情報通信学会パターン認識メディア理解研究会，福岡工業大学,PRMU2018-24，2018.9.20-21.

3. 劉家慶，健山智子，岩本祐太郎，陳延偉，“タッチベースインタラクティブ COVID-19 のセグメンテーション、定量評価 および可視化システム，” 電子情報通信学会医用画像研究会，MI2020-20，2020.9.3.

4. 劉家慶，黄越，黄辛隠，健山智子，岩本祐太郎，陳延偉，“センチメントテキスト朗読時の表情顔を用いたうつ状態の検出，” 電子情報通信学会パターン認識メディア理解研究会,PRMU2020-32，2020.10.9.

5. 劉家慶，黄越，黄辛隠，健山智子，岩本祐太郎，陳延偉，“CNN と Transformer エンコーダを用いたうつ状態の検出，” 電子情報通信学会パターン認識メディア理解研究会，PRMU2020-83，2021.3.

6. 劉家慶，柴樹榕，孫浩,黄辛隠,林蘭芬,健山智子,岩本祐太郎,陳延偉 Transformer エンコーダを用いたうつ状態の重症度の解析とマルチモーダルアダプティブアレーの融合に関する検討，” 電子情報通信学会医用画像研究会，信学技報，vol. 121，no. 98，MI2021-17，pp. 33-35，2021.7.9.

7. Wang Yi, **Liu Jiaqing**, Deng Zhuofu, Zhu Zhiliang, Chen Yen-Wei: “Development of a Collaborative and Mobile Platform for 3D Medical Image Analysis,” 第 36 回日本医用画像工学会大会，OP5-6，p.45，岐阜，2017.7.27-29.

8. 辻永成樹，山口展生，劉家慶，岩本祐太郎，健山智子，陳延偉：”L 字スクリーンと Kinect を用いた体感型 VR キャンパス案内システム，”平成 29 年電気関係学会関西連合大会，G12-4，近畿大学，2017.11.25-26.

9. 古澤康太郎，劉家慶，辻永成樹，健山智子，岩本祐太郎，陳延偉，“KINECT ハンドジェスチャ画像データベース構築と深層学習によるジェスチャ認識,” 平成 30 年電気関係学会関西連合大会，大阪工業大学，2018.12.1-2.

10. 木下将児，劉家慶，健山智子，岩本祐太郎，陳延偉，“Graph Convolutional Networks を用いた人体の 3 次元ポーズ認識” 映像情報メディア学会ヒューマンインフォメーション研究会，HI2021-8，2021.3.5.

Invited lecture

1. 劉家慶:“A Preliminary Study of Kinect-Based Real-Time Hand Gesture Interaction Systems for Touchless Visualizations of Hepatic Structures in Surgery”金森奨励賞の受賞記念講演、医用画像情報学令和 2 年度秋季(第 188 回)、オンライン、2020.10.3.

Award

1. 2016 年 10 月- 2018 年 9 月 文部科学省(MEXT) 国費外国人留学生奨学金
2. 2019 年 4 月- 2020 年 4 月 大塚敏美育英奨学財団 2019 年奨学生
3. 2020 年 4 月- 現在 日本学術振興会 DC2 特別研究員
4. 2017 年 9 月 13 日 国際学会 CMECE 2017 Best Student Presentation Award
5. 2018 年 11 月 11 日 国際学会 DMIP 2018 Best Student Presentation Award
6. 2018 年 11 月 7 日 立命館大学リサーチプロポーザルコンテスト大賞(最優秀賞で各分野 1 名のみ)
7. 2019 年 4 月 10 日 立命館大学大学院情報理工学研究科研究奨励賞
8. 2019 年 12 月 19 日 第 14 回「春暉杯」中国留学生クリエイティブベンチャーコンテスト 優秀賞
9. 2020 年 6 月 18 日 医用画像情報学会 金森奨励賞
10. 2019 年 6 月 立命館大学大学院博士課程後期課程 研究奨励奨学金 A
11. 2020 年 6 月 立命館大学大学院博士課程後期課程 研究奨励奨学金 S
12. 2021 年 6 月 立命館大学大学院博士課程後期課程 研究奨励奨学金 S

Acknowledgment

This thesis would not be possible without the support and guidance from many incredible mentors, labmates, friends, and family.

First and foremost, I am deep grateful to my supervisor, Professor Yen-Wei Chen, for his guidance, patient, and encouragement. His insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. Over the years, he guided me through all the hills and valleys in both research and life. I consider myself very fortunate for being able to study by his side throughout my graduate studies. Prof. Chen has been the best supervisor thar I can ever dream of. He will continue to be the best role model, both in academia and in life.

I would like to thank the dissertation committee: Professor Jien Kato, Professor Satoshi Tanaka for the thoughtful comments and recommendations on this dissertation.

I would also like to thank Professor Tomoko Tateyama from Shiga University and Dr. Yutaro Iwamoto from Ritsumeikan University for providing the in-depth discussion and helping me whenever being asked. I would also thanks Ms. Yuka Sato for the kind help in dealing with the documents.

In addition, I would like to thank Professor Lanfen Lin from Zhejiang University, China, Professor Xinyin Huang from Soochow University, China for their valuable suggestions and contribution to such good international cooperation.

Deep thanks go to all my co-authors, especially Mr. Ryoma Fujii, Mr. Kotaro Furusawa, Mr. Seiyu Tsujinaga, Mr. shurong Chai (Ristumeikan University, Japan), Ms Yue Huang (Soochow University, China), Mr. Hao Sun, Mr. Zhaolin Qiu and Ms. Huimin Huang (Zhejiang University, China) for their valuable comments and contribution to this research. I wish everything goes will in your future study and life.

I have also been very fortunate to be surrounded by many supportive labmates in Intelligent Image Processing Lab and all of my friends, they made my life at Ritsumeikan a truly memorable experience and their friendships are invaluable to me. I will miss our game nights, discussion, and many deadlines we fight together.

Next, this work is supported in part by Japan Society for Promotion of Science (JSPS) under Grant No. 20J13009 from 2020.4 to 2021.9. I would like to thank JSPS for generous financial assistance for this research.

I am grateful to Otsuka Toshimi scholarship Foundation. It is my honor to be selected as 2019 scholarship member. The assistance they provide give me a deeper understand of Japan and I will serve as a bridge between China and Japan as their expected.

I am grateful to Japanese government scholarship (MEXT) form 2016.10-2018.9.

Finally, I deep grateful to my family for their love and support all through my studies.

Jiaqing LIU
June 2021. Shiga, Japan