國立臺灣大學電機資訊學院電機工程學系

碩士論文

Graduate Institute of Electrical Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

通過深度視覺感知

對思覺失調患者的心理障礙檢測

Mental Disorder Detection for Schizophrenia Patients via Deep Visual Perception

林炳彰

Bing-Jhang Lin

指導教授：傅立成 博士

Advisor: Li-Chen Fu, Ph.D.

中華民國110年1月

January, 2021

1. 誌謝

1. 摘要

**關鍵字：**

1. ABSTRACT

**Keywords:**

1. TABLE OF CONTENTS

口試委員會審定書 #

[誌謝 I](#_Toc58427837)

[摘要 II](#_Toc58427838)

[ABSTRACT III](#_Toc58427839)

[TABLE OF CONTENTS IV](#_Toc58427840)

[LIST OF FIGURES V](#_Toc58427841)

[LIST OF TABLES VI](#_Toc58427842)

[Chapter 1 Introduction 1](#_Toc58427843)

[1.1 Motivation 1](#_Toc58427844)

[1.2 Lecture Review 3](#_Toc58427845)

[1.2.1 Emotion Recognition 3](#_Toc58427846)

[1.2.2 Depression Estimation 3](#_Toc58427847)

[1.2.3 Metric Learning 3](#_Toc58427848)

[1.3 Contributions 4](#_Toc58427849)

[1.4 Thesis Organization 4](#_Toc58427850)

[Chapter 2 Preliminaries 5](#_Toc58427851)

[2.1 Deep Neural Networks 5](#_Toc58427852)

[2.2 Metric Learning 5](#_Toc58427853)

[2.3 Facial Alignment 7](#_Toc58427854)

[REFERENCE 8](#_Toc58427855)

1. LIST OF FIGURES

Figure 2‑1: Visualized results of embedded features training on MNIST [3]. (a) softmax loss may result in the embedding with low intra-class compactness. (b) With the assistance of center loss, the trained model can encode samples well, thereby achieving higher feature discrimination. 7

1. LIST OF TABLES

# Introduction

In this chapter, we first describe the research motivation in Section 1.1. then elaborate on the lecture review in Section 1.2. After that, we concisely introduce the proposed algorithm and its contributions in Section 1.3 and conclude with the organization of this thesis in Section 1.4.

## Motivation

Schizophrenia is a mental disorder that progressively changes the mental state of a person and can be characterized by apparent changes in perception, thoughts, mood, and behavior. Even though people with schizophrenia often behave differently from someone else, their inner emotions and cognitive activities can be reflected through facial expressions and body language. Thus, emotion recognition becomes an essential technique for us to capture the deep insights of the patient.

In fact, emotion recognition is a pattern recognition task, and recent researches have almost focused on designing a neural network algorithm for learning robust embedding to yield discriminative features. Compared with traditional hand-crafted features, such as histogram of oriented gradient (HOG) and scale-invariant feature transform (SIFT), representations encoded via deep neural networks are more robust to motion and environmental changes, thus achieving better performance.

Previous emotion recognition strategies have been extensively discussed based on facial expressions. However, conventional facial expression recognition (FER) may fail to infer the real-time emotional state accurately. Due to facial muscle movements, such as blinking or opening the mouth, facial expressions may yield conflicting emotional signals, leading to incorrect and inconsistent predictions. On the other hand, facial expressions and body behaviors may represent different emotional signals simultaneously, such as facial expressions with surprise emotions *vs*. anger signals from defensive body language. In addition to the above, due to the nature of schizophrenia and the effects of drugs, patients particularly express fewer emotional signals, so that facial analysis alone may not be suitable for detecting the emotional state of patients. In cognitive science, some studies have shown that people recognize the emotions of others not only from their faces but also from the surrounding environment, such as the interaction of time series and the overall behavior of human appearance. Thus, integrating information from different forms is the key to achieving accurate recognition models. Moreover, certain symptoms of schizophrenia are often related to depression. Compared with patients with schizophrenia without depression, patients with depression have a worse treatment course and a worse prognosis.

We are thus motivated to design a multi-task learning framework to realize a mental disorder detection via both emotion recognition and depression estimation. To tackle the shortcomings of facial analysis and the inconsistency of different modalities (facial expressions and body behaviors), we design a cross-modality graph convolutional network (CMGCN), which can effectively determine a consistent signal from different modalities, consequently yields a robust representation. It is worth noting that our GMGCN takes advantage of the sparse sampling scheme to seek features with higher similarity and discard irrelevant elements, thereby achieving low computational cost and better model performance.

In addition to this effort, we also design objective functions for both tasks respectively to realize a better model convergence. For emotion recognition, which is a pattern recognition task, we propose a density loss with comprehensive criteria for metric learning relevant to orthogonal property and embedding density. By simultaneously suppressing inter-class interactions and encouraging intra-class connections, our density loss can constitute a robust embedding for inference.

In addition to the effort, we also propose a density loss (DL) with comprehensive criteria for metric learning relevant to orthogonal property and embedding density. By simultaneously suppressing inter-class interactions and encouraging intra-class connections, our density loss can constitute a robust embedding for identification. Compared with other state-of-the-art studies, our algorithm can achieve the most advanced performance with a clear margin.

## Lecture Review

We discuss the existing literature by emphasizing the following three most relevant aspects to our proposed method for mental disorder detection. Particularly, we first introduce the common emotion recognition techniques in Section 1.2.1, followed by depression estimation in Section 1.2.2. Finally, metric learning techniques are introduced in Section 1.2.3.

### Emotion Recognition

Emotion recognition is essentially a pattern recognition task and previous studies mainly focus on identifying human emotion based on facial analysis [1].

### Depression Recognition

### Metric Learning

The goal of metric learning is learning an embedding that projects the data from high dimensional to low dimensional space with high intra-class compactness and inter-class separability.

Since the redundant pairs bring drawbacks, less information and vain learning procedures, previous studies extensively explore to design mining and weighting schemes.

## Contributions

## Thesis Organization

# Preliminaries

In this chapter, some prerequisite knowledge is introduced. First of all, we give background information on deep neural networks, including convolutional neural networks and graph convolutional networks. After that, we discuss metric learning techniques for model convergence, including both *classwise* and *pairwise* scenarios. Finally, the facial alignment approach is introduced, which is adopted to localize the facial bounding box during the preprocessing stage.

## Deep Neural Networks

## Metric Learning

Metric learning is a way to encourage feature discrimination and is widely applied in many computer vision topics, such as pattern recognition, few-shot learning, and image retrieval tasks. It essentially focuses on learning an embedding to encode the same class data points to stay together and those of different classes to be faraway. Typically, its objective can be realized by carrying out a loss function to promote intra-class compactness and inter-class separability. According to the label assignments, metric learning can be built on two scenarios: classwise and pairwise. The former prefers to approximate the center of each class to realize the discriminative embedding, and the latter usually forms training data into pair or triplet relations and enforce a metric function to promote feature discrimination. In the following subsections, we introduce several classic metric learning techniques of both scenarios.

### Classwise Scenario

Softmax loss is the most widely used loss function for classification tasks and learns through the inner product among embedding features and the classification weight matrix for maximizing the probability of the target class. It can be expressed as:

|  |  |
| --- | --- |
|  | ( ‑ ) |

Where and are the batch size and the total number of classes, respectively. indicates a batch of embedding features and denotes the sample belonging to the class. denotes the classification weight matrix, that is the learned center of each class, and is the bias term.

However, since the nature of the inner product only instructs embedding features with a similar vector direction, softmax loss often results in sparse feature distribution. Thus, the learned embedding does not robust to the unseen data point. To alleviate this concern, center loss [2] is presented to further attract intra-class samples via an additional embedding center of each class and it can be expressed as:

|  |  |
| --- | --- |
|  | ( ‑ ) |

Where indicates an additional center embedding of each class.

Formally, center loss is an assistant loss term that further intensifies intra-class compactness, so it often works with softmax loss to realize a better intra-class compactness. Figure 2‑1 shows the visualization result between softmax loss and softmax loss with center loss. Compared with softmax loss, working with center loss can significantly improve feature discrimination, especially for class internal compactness, but we need to set the weights carefully. Since the center loss and softmax loss adopt different metrics for learning (Euclidean distance and inner product), we need to balance the range of loss values to avoid a futile learning procedure. Besides, the center loss usually requires extra effort in estimating the loss between each sample and its corresponding center; therefore, calculation and memory complexity must be considered.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
| Figure ‑: Visualized results of embedded features training on MNIST [3]. (a) softmax loss may result in the embedding with low intra-class compactness. (b) With the assistance of center loss, the trained model can encode samples well, thereby achieving higher feature discrimination. | |

Recent studies, which consider projecting features and classification weights into a bounded compactness sphere space, design various techniques by adopting different kinds of penalties to control the distribution of the embedding features, thereby resulting in a robust model. An angular softmax (A-softmax) [4] is proposed to map the features and the corresponding weights into the angular space. CosFace [5] and ArcFace [6] impose different margin penalty on the target weight for controlling intra-class compactness. As a matter of fact, these angular losses can be unified as a kind of sphere mapping, and it can be expressed as a general form by:

|  |  |
| --- | --- |
|  | ( ‑ ) |

Where the margin penalties of SphereFace [4], ArcFace [6] and CosFace [5] are respectively denoted as , and . For other notations, denotes the batch size and is the number of possible label classes. is the angle between the feature vector and its target weight vector and is the angle between the feature and other weight vectors.

Since the above sphere mapping techniques mainly focus on designing different penalties for intra-class perspective to realize the objective, inter-class separability is neglected. RegularFace [7] instead adopts an inter-class viewpoint for learning. It works by imposing a regularization term with the orthogonal property to regulate the similarity between inter-class weights. The regularization term can be expressed as:

|  |  |
| --- | --- |
|  | ( ‑ ) |

Where and denote the and classification weight, respectively. is the number of possible label classes.

However, this kind of regularization may lead to huge memory usage and ineffective learning procedure to large-scale datasets with large numbers of classes. Because the critical term is calculated from C×C cosine-similarity matrix, it may not suitable for large-scale classes.

Since the classwise scenario is limited to the class numbers of classes, it often calls for a rigid training procedure, not easily generalized to unseen classes. Even though classwise scenario takes the classwise viewpoint to learn a global guidance for each class

can provide better guidance for each class to realize better inter-class

## Facial Alignment

1. REFERENCE

[1] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing,* pp. 1-1, 2020, doi: 10.1109/TAFFC.2020.2981446.

[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in *Computer Vision – ECCV 2016*, Cham, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., 2016// 2016: Springer International Publishing, pp. 499-515.

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, vol. 86, no. 11, pp. 2278-2324, doi: 10.1109/5.726791.

[4] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 21-26 July 2017 2017, pp. 6738-6746, doi: 10.1109/CVPR.2017.713.

[5] H. Wang *et al.*, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18-23 June 2018 2018, pp. 5265-5274, doi: 10.1109/CVPR.2018.00552.

[6] J. Deng and S. Zafeririou, "ArcFace for Disguised Face Recognition," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 27-28 Oct. 2019 2019, pp. 485-493, doi: 10.1109/ICCVW.2019.00061.

[7] K. Zhao, J. Xu, and M. Cheng, "RegularFace: Deep Face Recognition via Exclusive Regularization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15-20 June 2019 2019, pp. 1136-1144, doi: 10.1109/CVPR.2019.00123.