Sentiment Analysis using Image,Text and Video

Chen Qian

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Doctor of Philosophy (Ph.D)

2022

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

28/July/2021

Chen Qian

. . . . . .

Date

. . . . . .

Chen Qian

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

28/July/2021

. . . . . .

Date

. . . . . .

Erik Cambria

Authorship Attribution Statement

This thesis contains material from 3 papers published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 3 is published as Chen Qian, I. Chaturvedi, S. Poria, E. Cambria and L. Malandri, ”Learning Visual Concepts in Images Using Temporal Convolutional Networks,” in IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp. 1280-1284, doi: 10.1109/SSCI.2018.8628703.

The contributions of the co-authors are as follows:

- Dr. Chaturvedi, Dr. Poria and I discussed the initial idea for the project and edit it.
- I collected the data, operated the experiments. A/Prof. Cambria provided guidance.
- I prepared the draft. Malandri assisted in the proofreading.

Chapter 4 is accepted as Chen Qian, E. Ragusa, I. Chaturvedi, E. Cambria and R. Zunino, ”Text-Image Sentiment Analysis,” accepted in CICLing (2018).

The contributions of the co-authors are as follows:

- Dr. Ragusa and I proposed the initial idea and discussed it with A/Prof. Cambria.
- Dr. Ragusa and I did the experiments and wrote the draft. Dr. Chaturvedi provided assistance and gave some suggestions.
- A/Prof. Cambria and Zunino helped revised the draft.

Chapter 5 is published as Qian Chen, I. Chaturvedi, S. Ji and E. Cambria, ”Sequential Fusion of Facial Appearance and Dynamics for Depression Recognition,” Pattern Recognition Letters for possible publication.

The contributions of the co-authors are as follows:

- I discussed the initial direction with A/Prof. Cambria.
- I collected data and conducted the experiments. Dr. Chaturvedi provided assistance in the experiments.
- A/Prof. Cambria and Dr. Ji revised the manuscript.

28/July/2021

Chen Qian

. . . . . .

Date

. . . . . .

Chen Qian

Abstract

Emotions and sentiments play a pivotal role in the modern society. In most human-centric environments, they are essential to assist decision-making, communication, and situation awareness. With the explosive increase in usage of social media (text, image and video) along with sentiment polarities for specific subjects (e.g., product reviews, political views and depression emotions), sentiment analysis has increasingly evolved as a subcomponent technology in lots of industries.

People are able to present their experience and feelings using images and there is a trend that people prefer image rather than just text. Compared with text, images provide more cues that better reflect people's sentiments and people can get a more perceptual intuition of sentiment. Particularly for the depression recognition problem in healthcare field, images containing human faces present emotions more intuitively with the face expressions. Hence, prediction of sentiment from visual cues is complementary to textual sentiment analysis. In this dissertation, studies are conducted to explore the sentiment analysis on media data ranging from image, image-text, to video data. We start from sentiment analysis on image data to explore the sentiment polarities. Then, investigations of sentiment analysis are conducted on images and their tags/captions, as such two types of data modalities provide more cues for improved sentiment analysis. Last, we explore the mystery of human emotions and dive into the issue of depression analysis on face videos. The main contributions of this thesis can be summarized as follows.

Firstly, for a single image, it may contain several concepts. To model the sequence of different sentiments of such concepts, we consider a Recurrent Neural Networks (RNN) besides Convolutional Neural Network (CNN). The proposed

Convolutional Recurrent Image Sentiment Classification (CRISC) model is able to analyze the sentiments of the context in one image without using the labels for the visual concepts.

Secondly, to explore the benefit of text data for image sentiment analysis, we propose to extract visual features by fine-tuning a 2D-CNN pre-trained on a large-scale image dataset and extract textual features using AffectiveSpace of English concepts. We propose a novel sentiment score to combine the image and text predictions and evaluate our model on the dataset of images with corresponding labels and captions. We show that accuracy by merging scores from text and image models is higher than using any one system alone.

Finally, we investigate multimodal facial depression representation by using facial dynamics and facial appearance. To mine the correlated and complementary depression patterns in multimodal learning, we consider a chained-fusion mechanism to jointly learn facial appearance and dynamics in a unified framework.

Therefore, this dissertation demonstrates our studies on image sentiment analysis, focusing particularly on facial depression recognition.

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor, Prof. E. Cambria. It is my great fortune to have the opportunity to pursue a Ph.D degree with his guidance and assistance. Without his support, I would not continue my study and come to the final.

Next, I would like to thank Dr. Iti Chaturvedi and Dr. Soujanya Poria for their inspiring discussions and assistance during my PhD study. Specially thanks to Dr. Haiyun Peng and Yukun Ma, who are senior PhD students of Prof. Cambria, for helping me to start the exploration of sentiment analysis.

This PhD was the most unforgettable experience of my life. I am really grateful to all my friends who cheered me up in difficult moments and shared my happiness when I made progress, including Dr. Pei Zhuang, Hui Li, Dr. Haiyan Yin, Dr. Wenya Wang, Dr. Zhu Sun, Xinghua Qu, Dr. Hanwei Qian, Dr. Jianjun Zhao, Dr. Longkai Huang, Yu Chen, Dr. Jing Guo, Dr. Yi Huang, Dr. Rui Yin, Mengshi Ge, Luyao Zhu, Wei Li, Tom Yang, Dr. Frank Xing, Dr. Sandro Cavallari, Dr. Yang Li, Shaoxiong Ji, Claudia Guerreiro, Dr. Edoardo Ragusa, Xin Zhao, Jiaqi Fan, Linlin Zeng and Erchen Song.

Finally yet importantly, I would like to thank my family members who always supporting me and encouraging me with their unselfish love.

Contents

List of Figures

List of Tables

List of Abbreviations

A-CNN	Appearance-Convolutional Neural Networks
AI	Artificial Intelligence
ANP	Adjective Noun Pairs
AVEC	Audio/Visual Emotion Challenge
BDI	Beck Depression Index
CD	Contrastive Divergence
CK	Cohn-Kanade
CNN	Convolutional Neural Networks
CRBM	Convolutional Restricted Boltzmann Machine
CRISC	Convolutional Recurrent Image Sentiment Classifier
DBM	Deep Boltzmann Machine
D-CNN	Dynamics-Convolutional Neural Networks
DM	Data Mining
EMFACS-7	Emotional Facial Action Coding System
FC	Fully-Connected
FER13	Facial Emotion Recognition 2013 Dataset
JAFFE	Japanese Female Facial Expression Dataset
JL	Johnson and Lindenstrausss
LPQ	Local Phase Quantization
MAE	Mean Absolute Error
MDD	Major Depressive Disorder
MHH	Motion History Histogram
MSE	Meam Square Error
NER	Named Entity Recognition
NLP	Natural Language Processing
PHOG	Pyramid of Histogram of Gradients
PLS	Partial Least Squares
RNN	Recurrent Neural Networks
RBM	Restricted Boltzmann Machine
RMSE	Root Mean Square Error
RP	Random Projection
SIFT	Scale-Invariant Feature Transform
SRHT	Subsampled Randomized Hadamard Transform

STIP	Space-Time Interest Point
SVD	Singular Value Decomposition
SVM	Support Vector Machines
SVR	Support Vector Regression
TISC	Text-Image Sentiment Classification
VSO	Visual Sentiment Ontology

List of Publications

International Journals

1. Qian Chen, Iti Chaturvedi, Shaoxiong Ji and Erik Cambria, "Sequential Fusion of Facial Appearance and Dynamics for Depression Recognition," *Pattern Recognition Letters*, 2021, vol. 150, p. 115-121.

2. Iti Chaturvedi, Qian Chen, Erik Cambria and Desmond McConnell, "Landmark Calibration for Facial Expressions and Fish Classification," *Signal, Image and Video Processing*, 2022, vol. 16, no 2, p. 377-384.

International Conferences

1. Luna Ansari, Shaoxiong Ji, Qian Chen and Erik Cambria, "Ensemble Hybrid Learning Methods for Automated Depression Detection," *IEEE Transactions on Computational Social Systems*, 2022.

2. Chen Qian, Iti Chaturvedi, Soujanya Poria, Erik Cambria and L. Malandri, "Learning Visual Concepts in Images Using Temporal Convolutional Networks," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, p. 1280-1284.

3. Chen Qian, Edoardo Ragusa, Iti Chaturvedi, Erik Cambria and Rodolfo Zunino, "Text-Image Sentiment Analysis," accepted in *CICLing (2018)*.

Chapter 1

Introduction

1.1 Background

When the first web appeared in 1989s, it was used for information-sharing among universities and institutes scientists in the world [2]. Since 2004, the proliferation of multimedia technologies and the increasing number of users communicating in virtual environment bring exponential growth of information online. From 2014, the third and also the current web came into view, which allows users the ability to interact with dynamic applications. Social media platforms such as FaceBook, Weibo, Twitter, and Tik-Tok are attracting millions of users sharing their daily life and opinions online. As reported in 2014, people send more than 500 million Tweets per day [1]. Valuable information may be hidden in those multifarious and innumerable data. Sentiment analysis aims to reveal the disposition of a person evoked when he/she meets a specific topic, person or entity [3]. As one of the most interesting and meaningful topic in Data Mining (DM) and Artificial Intelligence (AI), sentiment analysis attracts more and more researchers from both academic and industry and has numerous applications. For example, suppliers and sellers pine for the consumptive trend of their products and consumers make decisions according to product reviews simultaneously [4]. An study of the tweets political sentiment demonstrated that the tweets reflect the offline political landscape and can be served as a valid indicator to predict the political sentiment [5]. It also has been proved that the public mood are beneficial to the stock market prediction [6].

[1] https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html

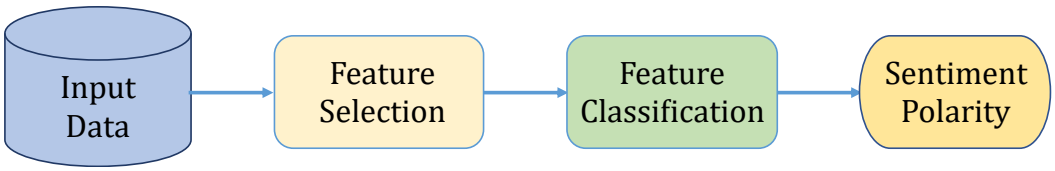


Fig. 1.1: The general process of Sentiment Analysis.

Further, studies on depression detection are playing an increasingly significant part in clinical diagnosis and monitoring [7, 8].

Essentially, sentiment analysis is a classification task and the general process is shown in Fig. 1.1. To identify the sentiment polarity of the given data, obtaining appropriate feature representations is the key which influences the accuracy of classification. Initially, sentiment analysis focused on text documents such as product reviews and comments posted on social media platforms (e.g., Twitter and Weibo). From the text, it was possible to extract the sentiment polarities of the sentences and classify them into positive, neutral and negative. It has been proposed [9] that sentiment classifiers for text can be trained from positive and unlabeled examples using machine learning techniques such as Naïve Bayesian method and Support Vector Machines (SVM). Afterwards, with the development and popularity of social media, millions of users are able and willing to share their views and lives in an offhand manner and in real time, which creates a huge amount of sentiment lexicons to be extracted and analyzed. Sentiment analysis methods based on sentiment lexicons appeared and neural networks enabled considerable progress in text sentiment classification [10–12]. Text-based sentiment analysis can be applied to various tasks, e.g., forex market prediction [13], movie polarity detection [14] and depression recognition [15]. It has been a long time that text-based sentiment analysis stood for sentiment analysis.

Apart from text, other modalities such as speech and vision came into consideration due to the popularization of multimedia leading to up-to-date expressions in various social networks. In this Big Data Era, image and video are novel ways of enabling human to express their thoughts and sentiments easily. Sentiment analysis is no longer limited to text domain. Particularly in the past few

years, studies on visual-based sentiment analysis have made considerable progress. Multiple large- and small-scale databases were constructed for sentiment analysis (e.g., [16–19]). Convolutional Neural Network (CNN) comes as the mainstream technique for image processing which can distinguish classes of images properly. In the works of [19–21], it has been proved that Deep CNNs are fitting for image sentiment analysis for its capability of detecting semantic concepts and learn high-level representations of them. CNNs not only show highly effective impact on images, but also manage useful facial signs for assessing emotions and depression in particular [22, 23].

1.2 Research Purpose

In this dissertation, our purpose is to improve the understanding of human affect and develop efficient and precise approaches to solve the image sentiment analysis problems especially depression recognition task. Besides the business use, sentiment analysis can be adopt in healthcare, e.g., clinical care of mental health disorders. As the most common mood disorder, depression is considered alone with persistent negative affect [24]. It has become a burning issue to recognize human emotions and detect depressive mood with significant value in both clinical and research applications. In clinical practice, there is a wide range of manifestations among distinct depressive disorders. Hence, the precise judgement of depression is essential for clinicians to conduct diagnosis and monitor the recovery process. We focus on the subtype named Major depressive disorder (MDD) and address the topic of depression over visual data. In the early phase, we began the image sentiment analysis on images from social media, and then shift to facial expression learning for depression recognition.

1.3 Challenges

The major challenges of this research are listed as follows:

Firstly, in the age of information exponential growth, explosive quantity of images are posted on social networks everyday, which means that all the objects in

our daily life may appear on the network. Currently, most of the machine learning approaches for sentiment analysis are designed for specific domains determined by labeled training data. However, label annotation for such large-scale dataset can be very costly, and hence sentiment analysis approaches without using the labels for visual concepts learning remains challenging.

Secondly, identifying the sentiment polarity of an image is not simply cumulating all the polarities of the image contents. The combination of different objects may leads to completely opposite sentiment polarities. To be specific, two faces with very slight variations may denote diametrically opposite emotions. In addition, some objects may be noises affecting the result of sentiment prediction. Feature extraction plays a principal role on sentiment analysis [25]. Thus, selecting beneficial information from abundant features is vital for improving the accuracy of image sentiment classification.

Thirdly, there are a lot of factors affecting the performance of sentiment analysis. Taking text-based sentiment analysis as example, puzzles such as negation, irony, ambiguous phrases and sarcasm exacerbate the challenge. For visual-based sentiment analysis, the within-cluster distance of images may be larger than the distance between different clusters. Such kind of cases often led to poor clustering of training instances in feature spaces [26].

Finally, multimodal learning seeks to build models that jointly exploit information from multiple modalities, and existing learning methods obtained significant advance by integrating different modalities at different stages. For multimodal sentiment analysis, there are several challenges, e.g., fusion strategy, hyper-parameter tuning, interpretability, and computational cost [27] when fusing the features of different modalities.

1.4 Contributions and Thesis Organization

In this dissertation, several approaches that aim to improve the feature learning of image sentiment analysis are illustrated. Specifically, our study covers three approach that work on image, image-text and video cues, respectively.

The major contribution of this research are listed as follows:

Firstly, we proposed an image sentiment analysis method termed Convolutional Recurrent Image Sentiment Classification (CRISC), which builds upon a temporal convolutional network and is able to analyze the sentiments of the context in images without using the labels for the visual concepts.

Secondly, we proposed a text-image sentiment classification (TISC) model to explore the benefit of text data for image sentiment analysis, where we propose to extract visual features by fine-tuning a 2D-CNN pre-trained on a large-scale image dataset and extract textual features using AffectiveSpace2 of English concepts.

Finally, we investigated multimodal depression analysis by using both facial dynamics and facial appearance. To mine the correlated and complementary visual depression patterns in multimodal learning, we introduce a chained-fusion mechanism to jointly learn facial appearance and dynamics in a unified framework. Our proposed multimodal fusion strategy differs from previous methods in its in-depth fusion both on feature-level and the prediction-level (score level), leading to boosted prediction performance.

The above mentioned approaches are presented in Chapter 3 to 5. This report is organized as follows:

- **Chapter 2 - Literature Review**

  In this chapter, we briefly introduce the evolution of sentiment analysis and image processing, including the background knowledge, problem statements, and the-state-of-art of research status.

- **Chapter 3 - Convolutional Recurrent Image Sentiment Classification**

  This chapter focuses on sentiment analysis over image database and covers the details of a framework for automatically learning visual concepts from images with the capability of identifying the sentiment of images as positive or negative. It is our first step on image sentiment analysis. Visual features representing sentiment concepts are extracted from a sequence of images from Flickr dataset and our proposed model promotes the performances significantly with the integration of deep convolutional learning and RNN.

- **Chapter 4 - Text-Image Sentiment Classification**

  To further probe into the feature extraction of visual sentiment, we observe the Twitter images with their captions for the prediction of sentiments. In this chapter, we conduct experiments on the dataset where the Twitter images have corresponding labels or tweets, hence the merging of features from images and text is proposed. In this way, we can predict image sentiment as positive or negative with better performance. We explored the model fusing text and image features is higher than using a single modality. To extract the image features we consider AlexNet, which is a pre-trained deep convolutional architecture. For text features we extract the significant concepts and project them on the AffectiveSpace of emotions. Lastly, we propose a novel sentiment scores to combine the prediction from image and text features.

- **Chapter 5 - Sequential Fusion of Facial Appearance and Dynamics**

  In this chapter, we explore the task of depression recognition over human faces. A deep multimodal learning method is proposed to learn the representation of facial appearance and dynamics. To model the correlated and complementary depression patterns between multi modalities, we propose a chained-fusion mechanism to jointly learn facial appearance and dynamics in a unified framework. Through the experiments, it has been proved that such sequential fusion provides a clear probabilistic perspective of the model correlation and complementarity between two different data modalities for improved depression recognition.

- **Chapter 6 - Conclusion and Future Work**

  This chapter summarizes our work and points out several directions required to be investigated in the future. In this dissertation, we study image sentiment analysis task, particularly visual depression recognition and achieve presentable results. We discuss the potential applications of multimodal depression recognition on textual-visual dataset to further promote the feature learning.

Chapter 2

Literature Review

2.1 Introduction

Since the birth of the Internet, it becomes easier to communicate with others wherever in the world. With the rapidly developing Internet, people are emerged one after another through social media platforms and their social communities have shown an increasing reliance on the network. Despite the Internet's role as a facilitator of information in this Big Data Era, it has overloaded users with unrelated and noisy data. It is essential to find an efficient and high-performance approach which extracts useful features from this massive amount of data. To accomplish this challenging task, researchers made efforts and obtained considerable achievements, but the time for satisfaction had not yet arrived.

In 2003, Nasukawa et al. defined the term Sentiment Analysis as *to find sentiment expressions for a given subject and determine the polarity of the sentiments* [28], while the sentiment analysis task was on text data as a subdomain of Natural Language Processing(NLP). Essentially, Sentiment Analysis is a classification process, where the output sentiment polarity can be binary, ternary or ordinary representation [29]. More specifically, the intensity of sentiment is investigated (e.g., in [22]). With the development of the technologies, sentiment analysis has cover not only text field but also multi-modal field and has thrived in recent years.

Initially, studies on sentiment analysis were focusing on text documents such as product reviews and comments posted on social media platforms. For text

data, there are certain semantic meanings for specific vocabularies/terms which makes it practicable to obtain the sentiment in a text. One application of textual sentiment analysis is preference prediction for products and movies. [30] studied the sentiment classes of movie reviews using Naïve Bayes, Maximum Entropy and SVM methods to categorize them into positive sentiment and negative sentiment. Afterwards, sentiment analysis methods based on sentiment lexicons appeared and neural network enabled considerable progress in text sentiment classification. It has been proved that sentiment polarities (like positive, neutral and negative) of text data are predictable and practicable for innumerable real-world applications which provides insights into activities of various industries and their customers such as tourism [31, 32], product recommendation [33], stock marketing [34] and so on.

Besides text data, image data also plays a vital role in sentiment analysis field and relevant research is attracting the attention of both academia and industry. In this context, the idea of image sentiment analysis emerged. While image involves semantic features with a much higher level of abstraction and human subjectivity [35], it is connaturally more challenging for visual sentiment analysis comparing with traditional recognition tasks. Predicting the sentiment of image requires a rich set of cues that leads to an exponential increase of calculation complexity. As it is difficult to model the sentiment due to the *affective gap* between the low-level visual features and high-level sentiment [36], it becomes more difficult than many other visual recognition tasks [37], e.g., object classification [38] and scene recognition [39]. One of the challenges for image sentiment analysis is the lack of benchmark datasets in this field. The main sources of data are from the posts on social media platforms and annotated manually. There might be disagreements on image labels as annotators often disagree between themselves, and even an individual is not always consistent with her/himself [40]. Properly annotated and large-scale image datasets are the foundation of image sentiment analysis. As a specific task of image sentiment analysis, depression recognition helps clinicians by predicting the value of a depression indicator and the indicator can serve as a guide to evaluate recovery. In current years, great efforts have

been put into building image databases with sentiment labels. Most of those databases are labeled as sentiment categories and converted to integers for the convenience in calculation process [41, 42]; some databases also use additional successive floats to represent the intensity of sentiment scores [19, 43]; and some databases provide integers as the intensity level of sentiments [17, 18]. Based on the databases, researchers started their journey on image sentiment analysis with traditional research disciplines such as Naive Bayes, Maximum Entropy, and SVM. The traditional algorithms can not meet the expectation on image sentiment analysis as there are innate limitations for them to analyze complex high-level sentiment features. In 1998, CNN was first founded and used in handwritten digit recognition [44] but did not attract much attention. Till 2012, Alex Krizhevsky proposed a powerful architecture for object recognition with CNN and showed the amazing ability to detect over 10,000 different objects simultaneously [45]. Afterwards, CNN becomes the mainstream technique for image processing and demonstrated impressive performance to automatically learn deep representations of images [46]. It also shows powerful capability on depression estimation by revealing possible correlation between depression and facial expressions [22, 47]. And it has been proved that deep convolutional networks were ideal for image sentiment analysis with the capability of extracting deep features of image [19–21].

2.2 Previous Work

In this section, we introduce the techniques and applications of sentiment analysis over various image databases. More specifically, emotion databases including depression dataset are introduced. For image sentiment analysis, we start from related works conducted over image datasets from social media. Researchers have started looking at ways in extracting visual sentiment features from images and attempts for depression recognition are conducted. Apart from visual-based sentiment analysis, here we first introduce related works for text-based sentiment analysis.

2.2.1 Text-based Sentiment Analysis

The proliferation of social media platforms and the increasing use of virtual communication resulted in exponential growth of information online. Despite the Internet's role as a facilitator of information in this Big Data Era, it has overloaded users with unrelated and noisy data. It is an urgent topic to find an efficient and high-performance approach which extracts useful features from this massive amount of data. NLP, a subdomain of AI, comes as a useful and practical method to handle the analysis of the language that humans use naturally in order to connect with computers and machines in both written and spoken contexts. For more than three decades, NLP has been handling problems between human and computer interaction.

NLP major tasks involve named entity recognition (NER), sentiment analysis, speech recognition, information retrieval, information extraction, relationship extraction, parsing, and machine translation, among others. Sentiment analysis, one of its most interesting and challenging tasks, combines advanced techniques from NLP, machine learning, and information retrieval to extract opinions and subjective knowledge from online messages in social media. In fact, the growth and broadening of social media enabled millions of users to share their opinions, hobbies and experience in an improvised manner and in real time, creating a huge amount of sentiment lexicons to be extracted and analyzed.

Initially, sentiment analysis focused on text documents such as product reviews and comments posted on social media platforms (e.g., Facebook, Twitter and Weibo). From the text, it was possible to extract the sentiment polarities of the sentences and classify them into positive, neutral and negative. It has been proposed that sentiment classifiers for text can be trained from positive and unlabeled examples using machine learning techniques [9] such as Naïve Bayesian method and SVM. Afterwards, sentiment analysis methods based on sentiment lexicons appeared and neural network enabled considerable progress in text sentiment classification.

As an innovative branch of NLP, sentiment analysis aiming to analyze the sentiment polarity of data, i.e., the attitudes, emotions and opinions of the data,

has attracted more and more interests and shown the benefits in various domains. It is widely used in our daily life, not only in the trivialities such as purchase decision, but also on the grand scale of global interactions, e.g., marketing decision and polls prediction. In early 1980s, Plutchik et al. proposed the idea named "Wheel of Emotions" which classify emotions as eight primary bipolar emotions and presenting emotions with a colorful wheel in terms of the connection between emotions and colors [48]. Plutchik's Wheel of Emotions serves as the foundation for sentiment analysis and make it more accessible to distinguish nuanced semantic concepts with intuitive sentiment presentation in term of visual perception. Cambria et al. proposed the hourglass model inspiring by Plutchik's studies [49]. Sentiments are roughly classified into three polarities: positive, neutral and negative. In the 3D hourglass model, affective states are shaped to an hourglass according to the sentiment polarity with regard to the strength of sentiment.

To explore users' behavior and opinion from numerous data, lexicons serve as the cornerstone of sentiment analysis by providing the theoretical basis for labeling and processing images. For text data, we list some of the most prominent sentiment lexicons in Section 2.3.1.

2.2.2 Image Sentiment Analysis

As sentiment analysis was born in the demand for text mining, consequently traditional studies on sentiment analysis are mainly appertaining to text data, while ignore data in other modalities, e.g., image and video data. Recently, image data becomes more and more popular and presents more information but with more complex in comparison to text data. The images extracted from online social platforms contain abundant objects with different scales. Image sentiment analysis is a challenging task. In this section, studies related to image sentiment analysis are illustrated.

In 2012, Siersdorfer et al. predicted sentiment of images using color histograms and Scale-Invariant Feature Transform (SIFT) techniques dataset with more than half a million Flickr images [50]. They used SentiWordNet as query terms to gather images with sentiment orientations. The bag-of-visual words representation and

the color distribution of images are used to learn the image features. Through studying the connection between sentiment of images expressed in metadata and their visual content, Siersdorfer et al. achieved the precision values of up to 70% but with low recall values. Zhang et al. integrated features from both text and image to process sentiment analysis on Microblogging [51]. In 2016, Katsurai et al. proposed a method mapping visual, textual and sentiment views into the latent embedding space and using correlations among these features [52]. The visual features is learnt from color histogram of images and this method achieved an accuracy of 74.77% on Flickr dataset and 73.60% on Instagram dataset.

As one of derivatives of the Plutchik's Wheel of Emotions, SentiBank [43] and DeepSentiBank [19] containing more than 3,000 Adjective Noun Pairs (ANPs) significantly improved in both annotation accuracy and retrieval performance [53], compared to its predecessors mainly using binary SVM classification techniques. Based on neural network, CNN introduced convonlutional filters to extract features and obtained outstanding achievements in image processing and deep learning. You et al. proposed a progressive CNN architecture [20] on CAFFE [54]. They trained half a million samples with ANPs from Flickr and fine-tuned the deep network using a progressive strategy therefore obtained a considerable accuracy with high recall. You et al. proposed to use CNN for the extraction of visual features and made fusions with textual features extracted from an unsupervised language model by learning distributed representations for documents and paragraphs [55]. Their model achieved a precision of 0.776 with recall of 0.740 by Early Fusion. Campos et al. provided a deep-dive analysis into *CaffeNet* and presented several experiments studying for the task of visual sentiment prediction [56]. In [57], Wu et al. proposed an approach combining global and local information that improved the performance of visual sentiment analysis. Salient object is detected to form local information and if there is no salient object detected, only global information is used for predicting sentiment. For emotion expression recognition, a subtask of sentiment analysis, Sanchez et al. imported self-attention mechanism to model the uncertainty in the temporal context of affective information and achieved significant improvement [58].

2.2.3 Depression Recognition

Generally, images containing human faces have sentiments stronger than natural scenes and are relatively easy to analyze. Extracting features from facial expression is a task having a history of more than forty years. The research topic of analyzing facial expressions has become a significant topic in machine vision study. Since 1970s, Ekman et al. [59] illustrated the universal facial expressions of emotion and studied six emotions: happiness, sadness, anger, fear, surprise and disgust. [60] found that facial action provides accurate information about a number of different aspects of the subjective experience of emotion, and the argument that even a type of smile may be a sign of negative emotion was still be made. Several decades ago, Friesen et al. created EMFACS-7 (Emotional facial action coding system) to do the initial facial emotion analysis [61]. During these decades, researchers never give up studies on emotions. Visual and audio data present a strong applicability on sentiment analysis. Van et al. analyzed physiology and mood through music by proposing a model trained based on skin and validated the concept of an affective music player directing the energy dimension of mood [62]. [63] detected emotions through not only facial expressions but also cries. [64] and [65] studied from facial expressions and audio. In researches of [66], an emotion space concept is used to process audio-visual emotion recognition. Gestures also show its significance in emotion recognition as [67] used gestures and speech information to do the emotion recognition task. [68] made a study of noise analysis in audio-visual emotion recognition. Furthermore, the large-scale in-the-wild facial expression dataset AffectNet was constructed and has promoted the development of emotion recognition.

Depression recognition is a subtask of emotion recognition as well as sentiment analysis in healthcare which could assist clinicians in the diagnosis and monitoring of depression. In mental health assessment, it is validated that nonverbal cues like facial expressions can be indicative of the depressive disorder. Recently, the multimodal fusion of facial appearance and dynamics based on convolutional neural networks has demonstrated encouraging performance in depression analysis. However, correlation and complementarity between different visual modalities

have not been well studied in prior methods. This section briefly reviewed two related topics: 1) visual-based depression recognition and 2) multimodal learning with deep architectures.

Automated Depression Recognition

Depression analysis based on various behavioral signals has drawn increasing attention in the affective computing community. Such feasible signals include the vision- and speech-based cues of human communication. While several works in the literature focus on this research topic, we are interested in the visual-based approaches for depression recognition. In the AVEC 2013 competition [17], a facial descriptor named the local phase quantization (LPQ) [69] was used as a baseline for facial depression recognition, where the extracted LPQ features for each video frame are further employed to train a support vector regression (SVR). In [7], Cummins et al. used the pyramid of histogram of gradients (PHOG) [70] and the space-time interest points (STIPs) [71] for extraction of behavioural cues for depression analysis. Meng et al. [72] proposed to use motion history histogram (MHH) [73] feature to model motion in videos, and then use the partial least squares (PLS) [74] for training regression model. In [23], the motion cue is encoded by the LPQ-TOP feature extracted from sub-volumes of the cropped facial regions, by which the behavior pattern dictionary can be obtained based on sparse coding.

In the AVEC 2014 competition [18], the local motion descriptor LGBP-TOP [75] and the SVR were employed as the baseline video description and prediction model, respectively. In [76], various local motion features extracted from sub-volumes of the detected faces were used for training an SVR-based prediction model. Jan et al. [77] proposed a 1D MHH based on some local descriptors to train a PLS regressor for final prediction. In [78], the baseline LGBP-TOP combined with LPQ was used as the video descriptor for depression prediction.

The aforementioned depression recognition methods proposed to use handcrafted descriptors, which are generally less effective to model and reveal high-level semantic cues. Recently, depression feature learning based on deep CNNs achieves

considerable progress. For example, Zhu et al. [47] proposed jointly learning the facial appearance and dynamics based on a two-stream CNN, in which two different features are fused at a fully-connected (FC) layer. Improved performance reported in their experiments indicated the efficacy of such a simple fusion manner. Most recently, Uddin et al. [79] introduced a new two-stream network for deep spatiotemporal feature learning, in which spatial information is extracted by a ResNet network, and they used a volume local directional number (VLDN) based feature descriptor to model facial motions. Zhou et al. [80] proposed a deep network named DepressNet to learn facial depression features with visual explanation, such that facial salient regions with different depression levels can be detected by using the generated activation mapping. Later, Zhou et al. [81] proposed to jointly learn the feature embedding and label distribution to address the issue of deep representation learning on a limited amount of labeled depression data, and the improvement by such learning scheme was reported in their experiments in comparison several state-of-the-art alternatives.

Multimodal Deep Representation Learning

In the multimodal setting, visual data consists of multiple input modalities [82–85], and each one may have a different representation and structure. Intuitively, useful representations could be learned from such multimodal data by fusing them into a joint representation to characterize the real-world semantic concept that the visual data corresponds to. In practice, however, it is much more difficult to model and discover the nonlinear correlation and diversity across modalities than those among features in the same modality. Recently, a good number of multimodal deep learning methods have been proposed to better exploit useful information from different modalities for more robust visual analysis [47,79,84,86–89]. In [90], Srivastava et al. proposed a multimodal learning method with a deep Boltzmann machine (DBM) to jointly learn multimodal feature representations. They approached this by adding a concatenated layer that connects DBMs from different modalities. In [87], a two-stream CNN with an additional multimodal fusion layer was proposed for RGB-D object recognition. Motivated by the observation that

the data from different modalities may contain modal-specific patterns as well as common patterns, Wang et al. [84] proposed a shareable and specific multimodal feature learning framework for RGB-D object recognition. By imposing the representation learning of associations between different modalities, Zolfaghari et al. [89] designed a chained multi-stream network to fully exploit the pose, motion, and appearance cues for action classification and detection.

Recently, multimodal representation learning has been well investigated in the field of affective computing. Yang et al. [91] proposed to combine the deep and shallow models for depression analysis using audio, video and text descriptors. In [92], Shang et al. also proposed to integrate advantages from hand-crafted and deep features. Differently, in their solution the facial images are encoded with local textures in the quaternion domain to retain the dependence of color channels. Niu et al. [93] proposed a multimodal attention feature Fusion strategy with spatio-temporal attention for multimodal depression representation. In [94], Mai et al. proposed to generate fused features via time-step level fusion at each time step, so that time-restricted interactions can be explicitly modeled by sharing information across modalities at the same time step. Despite the effectiveness of these multimodal learning approaches for different tasks, they typically consider the fusion either on the feature level or the prediction level. Very few attempts have been made on in-depth fusion that performs both on feature level and prediction level.

The aforementioned depression recognition methods proposed to use hand-crafted descriptors, which are generally less effective to model and reveal high-level semantic cues. As a result, these solutions typically lead to sub-optimal prediction performance.

Despite the success of deep learning based depression recognition methods reviewed above, fusion strategy of multimodal cues for effective depression prediction remains open. In particular, for video-based depression recognition, most previous methods perform multimodal fusion either on feature level or score level, without fully mining the potential of the multiple visual cues (i.e., appearance and dynamics).

Table 2.1: Semantic relations presented by pointers

	Antonymy	Hyponymy	Meronymy	Troponymy
Definition	opposing-name	sub-name	part-name	manner-name

2.3 Dataset

Emotional databases are indispensable resources and similar to textual sentiment lexicons in terms of visual domain. As multimedia becomes popular, visual and audio (even text-audio-visual) emotion databases and lexicons appeared. In this section, we list widely-used datasets for different image sentiment analysis domains with unimodal and multimodal data.

2.3.1 Text Databases

WordNet

WordNet is the earliest widely used lexicon for text mining. Miller proposed the idea of representing word senses by sets of synonyms (*synsets*) for machines to analysis text data [95]. WordNet containing more than 166,000 word pairs of form and sense (*f*,*s*) respects the syntactic categories *noun*, *verb*, *adjective* and *adverb*. Another contribution of WordNet is the following semantic relations represented by *pointers* between word forms or between synsets. Apart from *Synonymy*, there are four more relations shown in Table 2.1. With these semantic relations, systems for machine translation are able to achieve higher precision to determine which sense the author had in mind, which is helpful for information retrieval and opinion mining.

SentiWordNet

Based on WordNet, Esuli et al. created SentiWordNet by automatically annotating all the synsets of WordNet according to the notions of **positivity**, **negativity**, and **neutrality** [96]. To develop SentiWordNet, they conducted the quantitative analysis of the glosses associated to synsets, and then use the obtained representation vectors for semi-supervised synset classification. There are three numerical

scores for each synset s in WordNet, i.e., $Obj(s)$, $Pos(s)$ and $Neg(s)$, which describe the intensity of sentiment for the terms contained in the synset in terms of objective, positive, and negative. Each of the three scores are in range of 0.0 to 1.0 and with sum of 1.0 for each synset. There are numerous applications of SentiWordNet after its appearance. Denecke and Kerstin proposed an method which first translated other languages to English and then made multilingual sentiment analysis using SentiWordNet achieving accuracy of 66% in polarity classification [97]; SentiWordNet has proven a reliable resource for sentiment analysis in a multilingual context. Ohana et al. processed sentiment classification of movie reviews using SentiWordNet as a source of features for a supervised learning scheme pure term counting which indicated SentiWordNet could serve as an important resource for sentiment classification tasks [98].

WordNet-Affect

WordNet-Affect is a linguistic resource developed starting from WordNet. It provides lexical representation of affective knowledge containing 1,903 terms referring to mental [99]. Strapparava et al. manually realized a preliminary resource named AFFECT and identified the affective core by mapping the senses of terms in AFFECT to their respective synsets. After development of the *Core* of WordNet-Affect, an extension is made through WordNet relations. WordNet-Affect contains 2,874 synsets and 4,787 words and plays an important role in NLP tasks which need affective interaction [100, 101].

ConceptNet

ConceptNet is a exemplary Open Mind Common Sense corpus which was built from the vast unstructured information [102]. As the name of ConceptNet, it is a knowledge graph utilizing concept nodes to represent diverse words and phrases. The links between concept nodes provide more structural semantic information such as relations and attributes. One salient feature of ConceptNet is that it is able to automatically extracts conceptual facts from online data. Therefore, the number of concepts in ConceptNet will update over time. By providing accesses to

external knowledge, ConceptNet is applied in many NLP tasks like Commonsense Reasoning [103] and Question Answering [104].

SenticNet

Inspired by SentiWordNet, Cambria et al. developed a concept-level lexicon named SenticNet as a collection of polarity concepts which are known as common sense concepts and such common sense concepts are related to strong positive or negative polarity [105]. SenticNet is a semantic networks that links words and multi-word expressions to the emotion labels of the Hourglass model, i.e., Pleasantness, Attention, Sensitivity and Aptitude. Before SenticNet, most sentiment analysis resources were built by collecting the knowledge data from NLP resources such as WordNet or DBPedia and then manually annotated them. SenticNet make use of graph-mining and multi-dimensional scaling techniques on existing affective commonsense knowledge commonsense knowledge (e.g., WordNet-Affect, Open Mind Common Sense [106], and GECKA [107]) to automatically achieve the affective concepts. The latest SenticNet 6 is a much bigger knowledge base (of 200,000 commonsense concepts) than other sentiment lexicons.

AffectiveSpace

To build an appropriate knowledge base for the emotive reasoning task, Cambria et al. built AffectiveSpace based on the high-dimension concept vectors of SenticNet with a reduction technique. The resulted vector space contains continuous low-dimensional embeddings preserving the semantic and affective relatedness of the original space [108]. AffectiveSpace is improved by the Hourglass of Emotions and it is a vector space model containing 50,000 concepts with a 100-dimension vector for each concept, where the embeddings can be applied in potentially any real-world semantic tasks.

2.3.2 Unimodal Databases

This section illustrates unimodal databases for sentiment analysis available for common sentiment analysis tasks as well as emotion recognition and depression recognition.

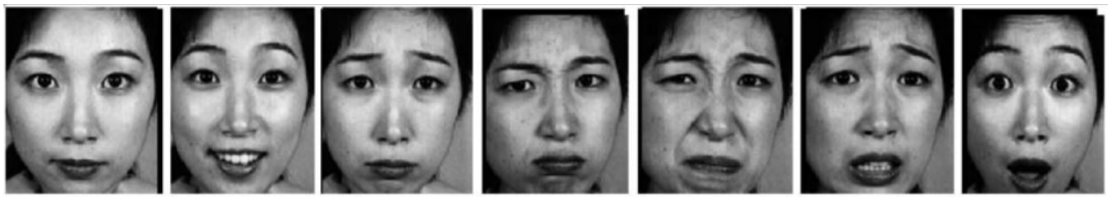


Fig. 2.1: JAFFE: facial expression images of neutral and six basic emotions

Fig. 2.2: CK+: facial expression images of seven basic and neutral emotions

Japanese Female Facial Expression Dataset

In 1998, Lyons et al. employed Japanese female expressers and subjects only and conducted the Japanese Female Facial Expression Dataset (JAFFE) to extract information about facial expressions [109]. The JAFFE dataset is a small-scale image dataset contains only 213 images of 7 facial expressions posed by 10 Japanese female models. Each image has been rated by 92 Japanese female undergraduates on a five-point Likert scale. JAFFE dataset is applied to emotion recognition as well as race recognition and gender classification [110]. Figure 2.1 shows six basic and neutral emotions from JAFFE dataset, namely *Neutral, Happiness, Sadness, Anger, Disgust, Fear and Surprise.*

CK/CK+ Database

Twenty years ago, aiming to promote research into automatically detecting individual facial expressions, the Cohn-Kanade (CK) database including 486 FACS-coded sequences from 97 subjects was released [111]. Since then, the CK database has become one of the most widely used test-beds for algorithm development and

(a) Anger (b) Disgust (c) Fear (d) Happiness (e) Sadness (f) Surprise (g) Neutral

Fig. 2.3: FER13: six basic and neutral emotions

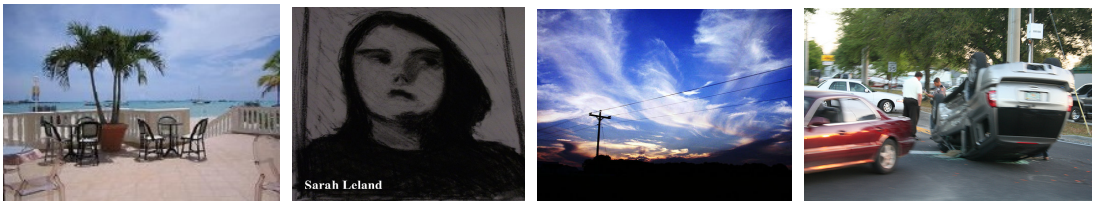
evaluation. In 2010, Lucey et al. extended the CK database to get a larger database CK(+) with 593 sequences from 123 subjects, and they added non-posed sequences for several types of smiles and their associated metadata [16]. Figure 2.2 are examples of the CK+ database and the emotions are *Disgust, Happy, Surprise, Fear, Angry, Contemp, Sadness and Neutral*.

Facial Emotion Recognition 2013

Facial Emotion Recognition 2013 (FER13) is an open-source data set that containing 35,887 gray-scale, 48x48 sized face images with seven emotions [41]. It is created by Pierre Luc Car-Brier and Aaron Courville. They searched images online with 184 emotion-related keywords using the Google image search API. Nearly 600 strings for facial image search queries were obtained by combining these keywords and words related to gender, age or ethnicity. So far, FER13 is one of the most widely-used dataset for emotion recognition and Figure 2.3 presents faces with six basic (*Anger, Disgust, Fear, Happiness, Sadness and Surprise*) and *Neutral* emotions from FER13 dataset.

SentiBank/DeepSentiBank

To handle image data as visual sentiment concepts, a novel system named Visual Sentiment Ontology (VSO) was demonstrated by combining sound structures from psychology and the folksonomy extracted from social multimedia [43]. Borth et al. proposed a pioneer concept, ANP, combines a noun for visual detectability and an adjective for sentiment modulation of the object described by noun semantics. VSO consists of 1,200 ANP concepts, such as happy dog and beautiful



(a) Beautiful View: 1.69 (b) Sad Face: -1.39 (c) Beautiful Sky: 2.0 (d) Bad Accident: -1.95

Fig. 2.4: SentiBank: ANP examples with sentiment scores

(a) Anger (b) Contempt (c) Disgust (d) Fear (e) Happiness (f) Sadness (g) Surprise (h) Neutral

Fig. 2.5: AffectNet: six basic and neutral emotions

sky, with manually labeled sentiment scores. To make full use of these ANPs, Borth et al. formulated a set of 1200 linear SVM outputs to represent sentiment concepts. These mid-level representations are called SentiBank.Complementary to this, there are abundant Flickr datasets containing more than 3,000 categories of image, each image belongs to one ANP and the number of images in one ANP is ranging from dozens to thousands. Figure 2.4 shows four examples for ANPs: "beautiful sky" and "beautiful view" reflecting strong positive sentiments (1.69 and 2.0 respectively) while "scary ghost" and "heavy smoking" reflecting strong negative sentiments (-1.39 and -1.95). In SentiBank database, the sentiment scores of ANPs are in [-2.0, 2.0]. On the basis of ANPs, different levels of different emotions may be extracted ([43, 55]). In [19], Chen et al. developed DeepSentiBank containing more than 3,000 ANPs significantly improved in both annotation accuracy and retrieval performance [20, 53].

AffectNet

AffectNet [22] dataset was developed by University of Denver which contains around 1M facial images collected from the internet. About 420K images have been manually annotated and the rest were automatically annotated using the

ResNext Network which had an accuracy of 65%. In our case, we used the manually annotated facial images as we wished to build a model on more accurate data. This dataset has 11 categories of images which includes categories like None, Uncertain and No-Face which isn't useful in our case. The other eight categories are *Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise* and *Neutral*. In AffectNet database, there are *Valence* and *Arousal* values assigned to each image. Valence value indicates how positive or negative an emotion is while Arousal value reflects whether an emotion is exciting/agitating or calm/soothing.

2.3.3 Multimodal Databases

In this section, multimodal databases involve audio-visual and EGG data are introduced.

eNTERFACE'05 Audio-Visual Emotion Database

The eNTERFACE'05 is an audio-visual emotion database contains 42 subjects with a two-week recording and it can be used as a reference database for testing and evaluating multi-modal emotion recognition algorithms [112]. The subjects were required to listen to 6 successive stories with a particular emotion and make reactions. There were two human experts judging the reaction and they added the reaction into the database when their got the same judgement of emotion expressed by subjects. In effect, eNTERFACE'05 was conducted to get and evaluate emotions as close as possible to spontaneous emotions.

IEMOCAP

Unlike eNTERFACE'05, Busso et al. designed a true-life interactive setting to elicit authentic emotions named the "interactive emotional dyadic motion capture database" (IEMOCAP). Specific types of emotions (happiness, anger, sadness, frustration and neutral state) were included in the database. IEMOCAP employed markers on the face, head and hands of the subjects. The recordings were from ten actors in dyadic sessions and their facial expressions and hand movements captured detailed emotional information during scripted and spontaneous spoken

communication scenarios [113]. There are approximately 12 hours of data in the corpus and the emotions of the subjects were genuinely expressed in a suitable context in addition to use plays (scripted sessions).

Spontaneous Filipino Emotion Database

Similar to JAFFE, Spontaneous Filipino Emotion Database is a emotion database collected within single race [114]. It focused on multimodal emotion recognition system that is trained using a spontaneous Filipino emotion database. Dy et al. collected 488 clips from the reality television show and the clips are ranging from 3 to 30 seconds. These clips were assigned to 20 Filipino students and classified into 5 emotions as happy, sad, angry, fear, and neutral.

DEAP

Distinct from traditional emotion analysis using visual features, Koelstra et al. proposed a database for emotion analysis using physiological signals, i.e., DEAP [115]. In the experiments, 32 subjects watched 40 one-minute music videos and their electroencephalogram (EEG) and peripheral physiological signals were recorded independently. Subjects rated each video The modalities of EEG and peripheral physiological signals were proved beneficial to explore the correlates between its frequencies and the participants ratings. For analysis of spontaneous emotions from physiological signals, DEAP has the highest number of participants in publicly available databases and the only database that uses music videos as emotional stimuli in 2012.

AVEC2013/2014 Database

Audio/Visual Emotion Challenge (AVEC) 2013&2014 [17,18] have the most widely-used depression subchallenge databases for depression recognition. In AVEC dataset, to evaluate the severity, a depression level score is measured by a self-reported 21 multiple choice inventory – Beck Depression Inventory-II (BDI-II) [116]. The BDI-II scores are in range of [0, 63], where the lower score represent more mild symptoms. The score-severity is shown in 2.2. For each video clip, there are 35 annotators predicting the BDI-II score.

Table 2.2: DBI-II scores and the corresponding severity of depression

BDI-II Score	[0, 13]	[14, 19]	[20, 28]	[29, 63]
Severity	minimal	mild	moderate	severe

In AVEC 2014, there are 84 subjects and each subject needs to perform two different tasks named Northwind and Freeform according to the instructions. All subjects in the two tasks speak German. There are 150 videos for each task and the data were split into three equal partitions: training, development, and test set. Each partition includes 50 videos and have similar distributions in terms of gender, age, and depression levels for the partitions. All videos are recorded by webcam in a human-computer interaction scenario 12265 and each video is approximately 2-minute length on average. There are at least 3 annotators per clip and most clips are annotated by 5.

Table 2.3: Notable Sentimental Databases

Dataset	**Size**	**Subject**	**Collected condition**	**Annotation content**
JAFFE [109]	213 images	10 Japanese females	Laboratory	The emotion is posed by the expresser
eNTERFACE'05 [112]	1166 video sequences	42 subjects coming from 14 different nationalities	Laboratory	Annotated by 2 human experts
IEMOCAP [113]	About 12 h videos	10 actors: 5 male and 5 female	Laboratory	Annotated by at least 3 human annotators
Spontaneous Filipino Emotion Database [114]	488 video clips	From a reality television show	Laboratory	20 Filipino students
CK / CK+ [16, 111]	486/593 video sequences	97/123 adults	Laboratory	Based on the subjects impression
DEAP [115]	1280 one-minute music videos	32 students	Laboratory	Self-assessment Manikins
FER13 [41]	35,887 images	Facial images searched by Google image search API	Laboratory	Emotion-related keywords
AVEC13/14 [17, 18]	300 videos	84 German speakers	Laboratory	A team of 5 naive raters
SentiBank / DeepSentiBank [19, 43]	About one million images	-	In the wild	ANP tags
AffectNet [22]	About 420K	-	In the wild	Annotated by 12 human experts

Chapter 3

Convolutional Recurrent Image Sentiment Classification [1]

3.1 Introduction

In this information exploring era, the Internet brings people rich information which also makes people under the pressure of data glut. To explore users' behavior and opinion from astronomical information, sentiment analysis came into being and has shown powerful capability on solving real-world problems. Traditionally, product reviews and comments posted on social media platforms (e.g., Facebook, Twitter and Weibo) were used for to develop sentiment analysis model. Each tweet was labeled as positive, neutral or negative using machine learning techniques such as Naïve Bayes and SVM [9].

With the rapid development of multimedia techniques in social network, text is no longer the sole mode for online communication. Tik-Tok short videos for example becomes a fashionable way to share lives especially among the young. Fig. 3.1 shows the examples of images with strong sentiment polarity as (a) is an image containing a happy baby laughing in the mother's arms; and (b) is an image showing an accident scene in ruins. Here, we try our first attempt of image sentiment analysis on Flickr images.

It is considered to extract visual sentiment concepts from images and has realized remarkable achievements, e.g., ANPs in SentiBank [43]. For video data,

[1] The content in this paper has been published in [117].

(a) Positive (b) Negative

Fig. 3.1: Examples for image polarities.

the tag accompanying the video is regularly employed to extract the concepts. However, tags may not exactly describe the images and sometimes even lead to opposite results. A ordinary way to deal with video data is to convert each video to a sequence of images and then analyze the sentiment in each image.

CNN shows its compelling capability in image precessing tasks, e.g., object recognition and classification. Coherently, attempts to investigate visual sentiment with deep Convolutional networks come into being. The results shows that deep CNN is ideal for image sentiment analysis as it can detect over 10,000 different objects simultaneously [19–21] as well as high-level semantic visual features. SentiBank is a database containing abundant images extracted from Flickr and the images are linked to more than 3000 ANPs with an adjective for sentiment modulation and a noun to describe the object. Based on ANPs, distinct emotions with different strength levels may be extracted [43,55]. In this chapter, we conduct experiments over the Flickr dataset from SentiBank. Generally, several visual concepts may occur in a single image, we consider utilizing additional Recurrent Neural Networks (RNN) to model them without using the ANP label [118].

In [119], Poria et al. blending CNN and RNN to identify sentiments of YouTube videos by investigating facial expressions. Inspired by their work, we propose the Convolutional Recurrent Image Sentiment Classifier (CRISC) model to handle the image sentiment analysis task on scenic images. Our model is evaluated over Flickr dataset and the experiment results outperformed baselines up-to 20%.

The organization of the chapter is as follows : in Section 3.2, we illustrate the background theoretical details of CNN and RNN. After that we detail the proposed framework of CRISC in Section 3.3. In the following section we evaluate the model on real-world Flickr dataset and lastly we summarize our contributions in Section 3.5.

3.2 Preliminaries

A deep neural network can be viewed as a composite of simple, unsupervised models. In our experiments, we take restricted Boltzmann machines (RBMs) and each RBM takes the output of hidden layer from the previous one serves as the visible layer. The gradient of the total energy function E in terms of the weights in all the layers is essential for the training of above multi-layer system. We maximize the global energy function with the approximate maximum likelihood contrastive divergence (CD) approach. The hidden state $\hat{h}_j$ is given by:

$$\hat{h}_j = b_j + \sum_i v_i w_{ij}, \tag{3.1}$$

where b_j is the bias of the jth hidden neuron, v_i is the ith visible node and w_{ij} is the connection weight to hidden neuron j from v_i.

The original deep neural network is consisted of several RBMs. To extended it to a convolutional deep neural network, the hidden layers are simply partitioned into Z groups [120]. A $n_x \times n_y$ filter is employed to each of the Z groups. Base on the assumption that the input image is of dimension $L_x \times L_y$, the convolution will generate a hidden layer of Z groups and each layer is of dimension $(L_x - n_x + 1) \times (L_y - n_y + 1)$. In a particular group, all hidden units are sharing the learned kernel weights. The energy function of layer l is given by:

$$E^l = -\sum_{z=1}^{Z} \sum_{i,j}^{(L_x-n_x+1),(L_y-n_y+1)} \sum_{r,s}^{n_x,n_y} v_{i+r-1,j+s-1} h_{ij}^z w_{rs}^l. \tag{3.2}$$

Therefore, each layer of a deep CNN is referred to as a convolutional RBM (CRBM). In such a model the lower layers learn abstract concepts and the higher layers learn complex features in terms of semantics. Lastly, due to the characteristic property of RNN in dealing with sequential data, it is employed to improve the learning speed and model the sequence of image features. The output at time step t for each layer of RNN ($\boldsymbol{x}_l(t)$) is calculated according to the following equation:

$$\boldsymbol{x}_l(t) = f(W_R^l.\boldsymbol{X}_l(t-1) + W_l.\boldsymbol{x}_{l-1}(t)) \tag{3.3}$$

where W_R is the interconnection matrix among hidden neurons, W_l is the weight matrix of connections between hidden neurons and the input nodes, $\boldsymbol{x}_{l-1}(t)$ represents the input vector at time step t from layer $l-1$, vectors $\boldsymbol{x}_l(t)$ and $\boldsymbol{x}_l(t-1)$ are hidden neuron activation at time steps t and $t-1$, respectively, and f is the non-linear activation function.

3.3 Proposed Framework

We proposed a novel approach by processing image data with a deep CNN and a low-dimensional RNN and the framework of our model is illustrated in Fig. 3.2. Image sequences are feed to the CNN for training to learn the visual sentiment features. To capture the temporal dependence, we transform each pair of consecutive images at t and $t+1$ into a single image. For the transformed input, two kernel with different dimensions are operated and denoted as Kernel 1 and 2 to learn 2D features of Layer 1. Layer 2 works in similar way as Layer 1. After that is a Upsampling layer to normalize the features with different kernel sizes. And there is a Logistic layer before RNN. With the delay states of RNN, our inter-connected hidden neurons can model long time delays which is helpful to accuracy improvement. Ultimately, each input image will be classified as "Positive" or "Negative".

CRISC Model

The proposed model is named *CNN and RNN Image Sentiment Classification* (CRISC) model which integrates RNN with CNN. The foremost procedure for

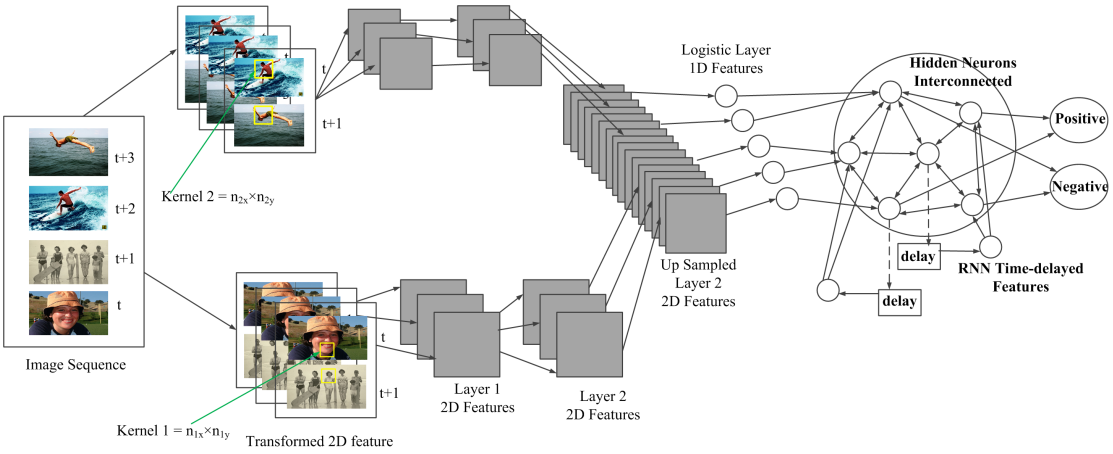


Fig. 3.2: Framework for CRISC

image analysis is to extract visual features from image properly. The feature extraction for the model is in two main steps.

We first construct a minimal deep CNN consisted of CRBMs where the visible layer has $L_X \times L_Y$ nodes. The first hidden convolution layer l of n_l features is of size $n_x \times n_y$. There are 5 kernels in each layer and in the first two convolution layers the dimension for kernels is 4×4. The reconstruction error is operated at every turn after adding a new layer. We keep adding layers till the error is the same as the previous one. Within each layer, we train the network with samples to learn the weights and construct new hidden layers of interconnected neurons when the reconstruction error is significant. Consequently, the expressions of training samples are modeled with learned elementary 2D features for image sentiment. For those features of different kernel sizes, we apply the Upsampling layer to transform them into uniform 2D features. Then, a logistic layer is employed before a RNN. Similarly, we construct RNN and learn weights in the same procedures as CNN, what different is the training data. Here, we have an inter-connected layer of neurons that can model long time delays using delay states. The final output layer serves as classifiers and gives each image a sentiment label.

3.4 Experiments

In this section we will introduce the feature extraction for image with sentiment, the classification results and evaluation of CRISC model and baselines for image sentiment analysis.

3.4.1 Datasets

In this work, we conduct the experiments on a Flickr[2] image dataset with IDs and sentiment annotations. Comparing to the original dataset, some images are unavailable for security reason. There are thousands of images annotated by three different annotators as three sentiment polarities (positive, negative and neutral). In our experiments, we removed neutral images and then get 19,255 positive images and 7,526 negative images. The input of our algorithm is a sequence of images with sentiment labels. To save the computing cost, we reduce the resolution of the input images to a size of 30×30 and split the dataset into 9 sets (8 sets of 3,000 images and one of 2,781) for experiments. Each of the 9 image sets is divided into three partitions: 2,000 training images, 500 validation images and the rest are testing images.

3.4.2 Experiment Results

Table 3.1 shows the results of four sets with different parameters. In the experiments, we test the results with the setting of (a) neu = 20, epo = 50; (b) neu = 10, epo = 50; (c) neu = 5, epo = 100. Where *neu* is the number of neurons and *epo* denotes the training epochs. The variance across 9 subsets of Flicker data is high. To evaluate our approach, we compare the classification accuracy of our approach and baselines with the accuracy defined as:

$$Acc = (tp + tn)/(tp + fp + tn + fn) \tag{3.4}$$

[2]http://mm.doshisha.ac.jp/en/senti/CrossSentiment.html

Table 3.1: Accuracy of CRISC in different parameters

No.	**(a)**	**(b)**	**(c)**	**Ave**
(0)	65.0%	65.0%	65.5%	65.0%
(1)	88.0%	**96.0%**	90.0%	**91.3%**
(2)	74.5%	75.5%	75.5%	75.2%
(3)	75.5%	76.0%	77.5%	76.3%
(4)	68.5%	68.5%	68.5%	68.5%
(5)	71.0%	70.5%	70.5%	70.7%
(6)	79.5%	79.5%	79.5%	79.5%
(7)	79.5%	79.5%	79.5%	79.5%
(8)	65.1%	68.7%	68.7%	67.5%
Ave	74.1%	**75.6%**	75.0%	74.9%

3.4.3 Discussion

As shown in Table 3.2, the accuracy of random method for binary classification is about 50%; when using SentiBank, a mid-level visual feature-based method, the classification accuracy achieves 70.01%; by operating SentiBank with a low-level visual feature-based method [50], Low&SentiBank [121] achieves similar accuracy as SentiBank (about 70.54%); in the methods from forth to seventh [52], extra views are conducted to calculating latent correlations, which are denoted by LC, and the features projected from images for classification are shown by **P** with Text (**T**), Visual (**V**) and Sentiment (**S**) features, the accuracy of LC-based method is in range of (64.63%,74.77%); CNN built upon AlexNet has accuracy of 76.38% and our method achieves 78.26%. The results shows that our approach is significantly outperforming other approaches, in the range of [65.2- 91.33]%. Hence, we outperformed baselines up-to 20%.

3.4.4 Evaluation

It is not universal to solely study in image data. The mainstream is to joint image and text or some other views on image. We conduct the experiments on image data and our model is highly dependent with the dataset. Subset 1 (out of 9) achieved the highest accuracy of 96% while the accuracy declines slightly for other subsets.

Table 3.2: Comparison with Baselines on Flicker Data

Method	**Accuracy**
Random	49.78 ± 1.05%
SentiBank [50]	70.01 ± 0.63%
Low&SentiBank [121]	70.54 ± 1.00%
LC(T+S)+P(T+S)	64.63 ± 0.91%
LC(V+S)+P(V)	70.67 ± 0.78%
LC(V+S)+P(V+S)	68.98 ± 1.01%
LC(V+T+S)+P(V+T+S)	74.77 ± 0.82%
CNN(AlexNet)	76.38 ± 0.79%
CRISC	**78.26 ± 13.07%**

3.5 Conclusions

In this chapter, we have proposed a framework to automatically learn the visual concepts and classify the sentiment of images. The proposed method works without ANP tags from web where there contains a lot of noise. Experiments on Flickr benchmark shows that our model outperforms baselines significantly. We extract visual concept features from a sequence of images by deep convolutional learning and then feed the visual concepts in a single image to RNN. The lower layers of the deep neural network consisting of several CRBMs can learn abstract features and then these features are combined to form representations of visual sentiment concepts in the higher layers.

Chapter 4

Text-Image Sentiment Classification [1]

4.1 Introduction

Since the popularization of multimedia content in various social networks, text is no longer the only mode for sharing information. Image and videos are enabling people to express their thoughts and sentiments easily. As a consequence in this Big Data Era, sentiment analysis cannot be limited to text domain. In particular, images play a more important role in sentiment analysis, since they have the fullest quality of information. Furthermore, videos can also be represented as a sequence of images. Expressiveness varies from one person to another. Most images posted on Twitter lack good labels and the accompanying tweets have a lot of noise. For example, YouTube videos are a convenient way to share news events and product descriptions. In Figure 4.1, (a) illustrates an image of two gentlemen paddling their canoes and laughing, which is annotated as positive polarity; and (b) is an image of a building in ruins annotated as negative polarity.

CNN are the mainstream technique for image processing which can distinguish classes of images properly. Several authors have used CNN for object recognition and classification of images [19, 20, 45, 55]. It has been proved that Deep Convolutional networks were ideal for image sentiment analysis as it is able to distinguish more than 10,000 different objects. This powerful CNN architecture named

[1] The content in this paper has been published in [122].

(a) Positive

(b) Negative

Fig. 4.1: Examples for image polarities.

AlexNet [45] is used in our method. Complementary to this, there are abundant Flickr datasets containing more than 3,000 ANPs, where each image belongs to one ANP and the number of images in one ANP is ranging from dozens to t-housands [19]. On the basis of ANPs, emotions may be detected along with the intensity of sentiment [43, 55].

In this chapter, we propose a method using textual and visual features to predict the sentiment polarity of Tweets containing both image and text. The Twitter dataset contains images for visual feature learning. Following is the structure of this chapter: Section 4.2 is the preliminaries for our approach; Section 4.3 states our methods in a detailed way; Section 4.4 is the results and evaluation for our methods; finally, we summarize our work in Section 4.5.

4.2 Preliminaries

In this section, we will give the theoretical basis about CNN and AffectiveSpace 2 for our method.

4.2.1 Deep Convolutional Neural Network

CNN are a specific class of neural networks, based on four main building blocks: convolutions (kernels), non-linearities, pooling and dropout layers. A CNN is comprised of one or more convolutional layers (kernels) alternated by non linearities.

Between them are inserted pooling and dropout layers. Finally, one or more fully connected layers as in a standard neural network gives the classification results.

Each convolutional layer works as a feature extractor. Using objects recognition as an example, lower level detects simple features like straight edges, simple colors, and curves, higher level extracts more complex features like noses, eyes. Typical non linearities are *ReLu* and *Tanh*. Dropout layer are a regularization technique for reducing over-fitting in neural networks by preventing complex co-adaptation on training data. Max pooling layers performs down-sampling by dividing the input into pooling regions, and computing the maximum of each region. The fully connected layer merges the extracted feature in order to perform the classification task.

This models present a huge number of parameters and are trained using standard back propagation techniques. Training from scratch requires labeled datasets with millions of patterns. For this reason in many applications transfer learning is applied. Transfer learning consist in remove the last fully-connected layer from a fully trained CNN, and replacing it with a new one. Then fine-tuning is applied to the weights of the new network by continuing the back -propagation. The main advantage of this technique is that it is possible to exploit feature detector for similar tasks, as an example adapt features for object recognition to polarity detection.

4.2.2 AffectiveSpace 2

To improve our model with textual features, we projected textual tweets data to AffectiveSpace 2. It is an effective way to cope with the evergrowing number of concepts and semantic features using AffectiveSpace 2. Cambria et al. replaced Singular Value Decomposition (SVD), a low-rank approximation method, with random projection (RP) [123] to map the original high-dimension features into a much lower-dimensional subspace without losing the semantic and affective relatedness information via Gaussian matrices, hence the pair-wise distances were maintained with high probability. This follows Johnson and Lindenstrausss (JL)

Lemma [124]. The JL Lemma states that with high probability, for all pairs of points $x, y \in X$ simultaneously, there is:

$$\sqrt{\frac{m}{d}}\|x-y\|_2\,(1-\varepsilon) \leq \|\Phi_x - \Phi_y\|_2 \leq \sqrt{\frac{m}{d}}\,\|x-y\|_2\,(1+\varepsilon) \tag{4.1}$$

where X is a set of vectors in Euclidean space, d is the original dimension of this Euclidean space, m is the dimension of the space we wish to reduce the data points to, ε is a tolerance parameter measuring to what extent is the maximum allowed distortion rate of the metric space, and Φ is a random matrix.

Sarlos introduced that structured random projection for making matrix multiplication much faster [125]. Achlioptas and Dimitris proposed sparse random projection [126] to replace the Gaussian matrix with i.i.d. entries in:

$$\phi_{ji} = \sqrt{s}\begin{cases} 1 & with\ prab \cdot \dfrac{1}{2s} \\ 0 & with\ prob.1 - \dfrac{1}{s} \\ -1 & with\ prob.\dfrac{1}{2s} \end{cases} \tag{4.2}$$

where one can achieve $a \times 3$ speed up by setting $s = 3$, since only one third of the data need to be processed.

Subsampled randomized Hadamard transform (SRHT) is preferred to manage data when the number of features is much larger than the number of training samples ($d \gg n$). The way SRHT works is extremely approximate to Gaussian random matrices but with lower Computational complexity ($O(nd)$) than Gaussian matrices ($O(nlogd)$) [127]. From [127,128], for $d = 2^p$ where p is any positive integer, a SRHT can be defined as:

$$\phi = \sqrt{\frac{d}{m}}RHD \tag{4.3}$$

where d is the number of features and m is the customized number to randomly subsample from the whole features; R is a random matrix with $m \times d$ size (the rows of R are m uniform samples from the standard basis of $\mathbb{R}^d$); $H \in \mathbb{R}^{d \times d}$ is a normalized Walsh-Hadamard matrix defined recursively as follows:

$$H_d = \begin{bmatrix} H_{k/2} & H_{k/2} \\ H_{k/2} & H_{k/2} \end{bmatrix}, \tag{4.4}$$

$$H_2 = \begin{bmatrix} +1 & +1 \\ +1 & 1 \end{bmatrix} \tag{4.5}$$

and D is a $d \times d$ diagonal matrix where the diagonal elements are i.i.d. Rademacher distributed random variables.

AffectiveSpace 2 is a vector space model preserving the semantic and affective relatedness of common-sense concepts while being highly scalable [108]. In our method, it is an important part to extract sentiment features from textual Twitter data using AffectiveSpace 2.

4.3 Proposed Framework

This chapter proposed approach consists in merging the information from images and their captions. The feature from the image and the caption are extracted independently. The text features are extracted by means of single world projections on AffectiveSpace 2. The visual feature are extracted using AlexNet and fine-tuned by Twitter images.

4.3.1 Visual features

The proposed feature extraction model of the CNN is inspired from AlexNet. Since AlexNet is a deep CNN architecture trained on 1.2 million images for the task of object detection with considerable precisions, it is efficient to detect sentiment of images by fine-tuning AlexNet with labeled images. In our model, we removed the fully connected layers and replaced it with a fully connected layer of size 4096×2. Then we fine tune the weights using about 18,928 pattern from Twitter, derived from 301 negative images and 581 positive images.

4.3.2 Textual features

The textual features are extracted by 5-fold cross-validation method with AffectiveSpace. The original samples are randomly partitioned into 5 equal sized subsamples. Of the 5 subsamples, we train four of them and the other one is retained as the validation data. This process is repeated 5 times. SVM is used in the procedure of extracting textual features.

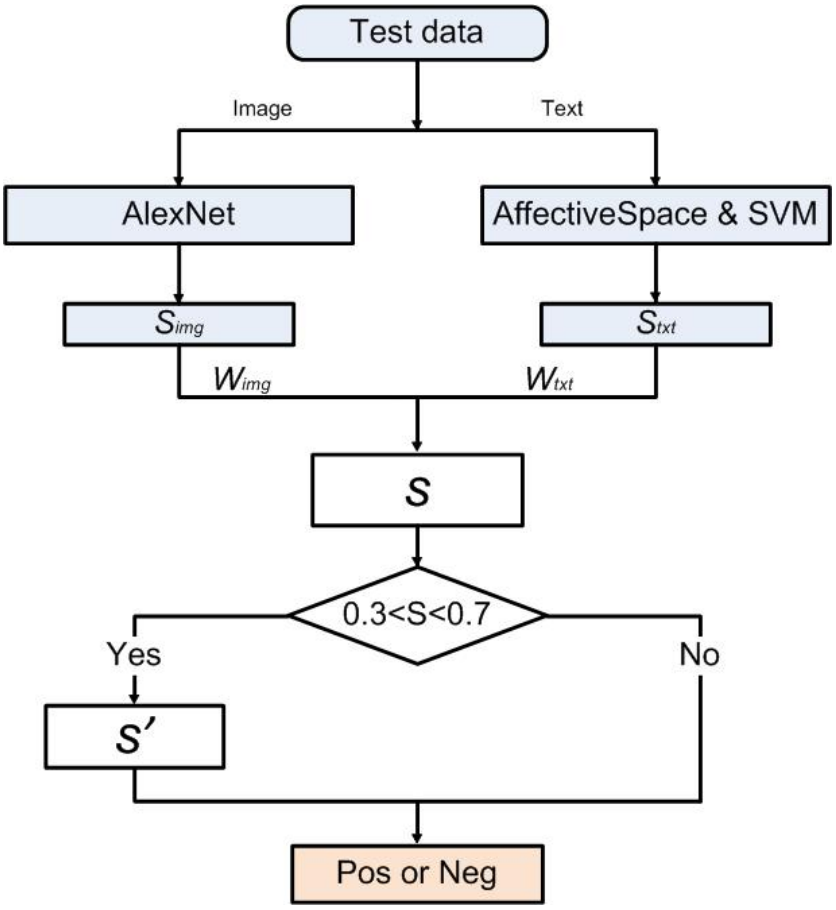


Fig. 4.2: Framework for TISC

4.3.3 TISC Model

Sentiment polarity prediction is based on features extracted from images and texts respectively. In order to balance visual features and textual features, our Text-Image Sentiment Classification (TISC) model employs the following computational approach to obtain sentiment scores of test data. Fig. 4.2 is the flowchart of computing sentiment scores. First of all, we define the preliminary weight of our image features extraction architecture as:

$$w_{img} = \frac{Acc_{img}}{Acc_{img} + Acc_{txt}} \tag{4.6}$$

where Acc_{img} and Acc_{txt} is the accuracy for validation of image and text data respectively. Similarly, the textual weight is $w_{txt} = 1 - w_{img}$.

The preliminary sentiment score s is calculated by the following equation.

$$s = s_{img} w_{img} + s_{txt} w_{txt} \tag{4.7}$$

where s_{img} is the sentiment score for test image predicted by CNN, while s_{txt} is the score for text data given from AffectiveSpace 2.

For test data with $s \in [0.3, 0.7]$, we assume that such data do not have strong sentiment polarities or there is conflict between textual and visual features. Hence, using the weighted scores to classify the sentiment polarities of these data is not appropriate. We import a new measure s' with variables $\alpha, \beta \in [0, 1]$ giving weights to visual and textual system respectively:

$$s' = 1 - (\alpha s_{img} + \beta s_{txt}) \tag{4.8}$$

With the sentiment scores measured as above, we define the sentiment polarity with 1 for *Positive* and 0 for *Negative* calculated by the following equation:

$$Polarity = \begin{cases} 1 & with\ s \geq \alpha \ \ or \ \ s' \geq 0.5 \\ 0 & with\ s \leq \beta \ \ or \ \ s' < 0.5 \end{cases} \tag{4.9}$$

where α and β are experimental parameters which we set to 0.7 and 0.3 respectively.

4.4 Experiments and evaluation

In this section we first introduce the feature extraction from images for sentiment classification. Next, we compare the accuracy of TISC model with baselines for image sentiment analysis.

4.4.1 Datasets

To conduct experiments on textual and visual features, we fine-tune AlexNet with 1,269 labeled Twitter images[1]. These images are annotated by 5 Amazon Mechanical Turk workers and we choose the images with the same sentiment label

[1] http://www.cs.rochester.edu/u/qyou/DeepSent/deepsentiment.html

Table 4.1: Results of different methods

Method	α	β	Pred_neg	Pred_pos	Rec_neg	Rec_pos	Accuracy
Visual feature	-	-	0.3878	0.8352	0.4385	0.8043	0.7169
Textual feature	-	-	0.4122	0.8439	0.4692	0.8109	0.7356
TISC(1)	0.054	0.448	**0.6596**	0.8177	0.2385	0.9652	**0.8051**
TISC(2)	0.284	0.244	0.5574	0.8185	0.2615	0.9413	0.7915
TISC(3)	0.030	0.578	0.4787	0.8286	0.3462	0.8935	0.7729
TISC(4)	0.300	0.340	0.4058	0.8371	0.4308	0.8217	0.7356
TISC(5)	0.792	0.798	0.3610	0.8768	**0.6692**	0.6652	0.6661

given by all the 5 workers (581 positive and 301 negative images). Since the data are unbalanced, in order to improve the fine-tuning on AlexNet, we double the negative images and increase the size of dataset to 18,928 by adding rotations and reversals of each image. To judge our model, we test it on 596 images (463 positive and 133 negative images) with captions from Twitter [121].

4.4.2 Experiment Results

In the first step of our approach, we fine tune the AlexNet with Flickr dataset to obtain the visual kernel features. Next, textual features are extracted by 5-fold validation using SVM classifier. To find the optimal combination of textual features and visual features, we use trial and error method, we progressively change all the parameters $\alpha, \beta \in [0, 1]$ in step of 0.002 (251,001 pairs of α and β). The highest accuracy in Fig. 4.3 reached 0.8051 and the accuracy is stable in the left part. Fig. 4.4 shows how the value of α or β effects the accuracy and for each curve, the accuracy of α is the average accuracy for β on each value of α, which is similar to β. Table 4.1 shows the results of sentiment prediction using different methods and TISC using diverse parameters.

Table 4.2: Comparison of other methods evaluated by AUC

Method	AUC
Low-level Features [121]	0.528
SentiBank [43]	0.514
TISC(VGG+TFIDF)	0.547
TISC	**0.586**

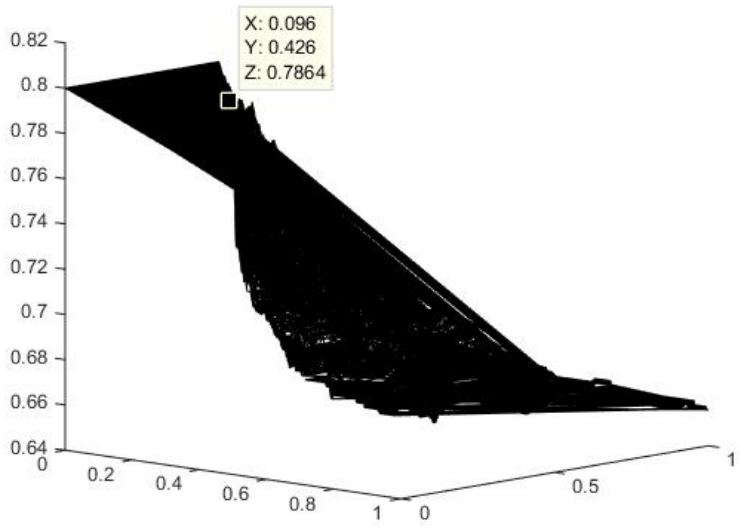


Fig. 4.3: Accuracy of Text-Image

4.4.3 Evaluation

Fig. 4.4 shows the accuracy of TISC corresponding to different α and β. For example, the accuracy achieves 0.7864 at point $(0.096, 0.426)$. From Fig. 4.3, it is shown that for α, β, if $\alpha + \beta \in (0, 0.5)$, then the results of system are stable with accuracy of 0.8051. However, from eqn (4.8) reveals the truth that a high accuracy is with a low recall for Negative data since when $\alpha + \beta < 0.5$, for all data there is $s' > 0.5$ and then be classified as positive data. Fig. 4.4 is the trend curves for accuracy-α/β from where we can see that in [0,1], the higher the values of α/β is, the lower the accuracy is.

From Table. 4.1, the last five rows are the results reflecting to different (α, β) pairs with $\alpha + \beta > 0.5$, we can find that the TISC has a significant increase in sentiment prediction compared with using single measures to predict sentiment polarity of test data and precision and recall are negative correlated. Table 4.2 shows that our model outperforms baselines by almost 6% in AUC. We compare with two baselines: Low-level Features are a set of features that can be useful for characterizing sentiment clues such as scenes, textures, and faces as well as

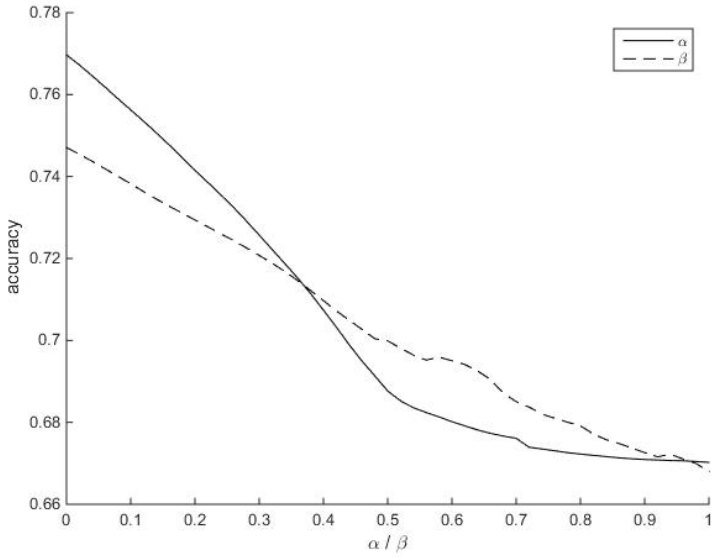


Fig. 4.4: Accuracy -α/β curve

other abstract concepts [121]; SentiBank is a concept representation with detectors trained on Flickr images.

4.5 Conclusion

This chapter considers the application of Twitter images with captions for the prediction of sentiments applying fine-tune techniques. The Twitter images have corresponding labels or tweets, hence the merging of features from images and text is proposed. In this way, we can predict image sentiment as positive or negative with better performance. We see that the accuracy after fusing text and image features is higher than using a single modality. To extract the image features we consider AlexNet, which is a previously trained deep convolutional architecture. For text features we extract the significant concepts and project them on the AffectiveSpace of emotions. Lastly, we propose a novel sentiment score to combine the prediction from image and text features.

Chapter 5

Video-based Facial Depression Analysis [1]

5.1 Introduction

In the above chapters, we explore the image sentiment analysis task with general labels for negative sentiments over datasets from social media which is far from satisfaction for depression recognition. As the depressive sentiments are always along with intensities, approaches more sensitive to slight variations are needed in depression recognition tasks. MDD is a psychological disorder that exhibits feelings of sadness, loss, or anger that may impact a person's usual social activities. At a global level, over 300 million people of all ages suffer from different levels of depression, equivalent to 4.4% of the world's population [24]. A depressive episode can be classified into a minimal, mild, moderate, or severe level, depending on the symptoms. Mild depression may bring difficulties in continuing with ordinary work and social activities. More seriously, the feeling of depression may occur comorbidity with self-mutilation [129], and depressed people are more likely to commit suicide than the general population [130]. Early detection of depressive or other mental disorders provides a possible way for mental intervention [15].

In clinical practice, the diagnosis procedure for MDD can usually be labor-intensive and highly relies on expertise observations. Due to the increasing number of people suffering from MDD around the world, methods for automated

[1] The content in this paper has been published in [1].

depression analysis appear to be urgent for objective and efficient diagnosis. Recently, automated depression diagnosis based on computer vision techniques has drawn increasing attention [18], and the significance of the verbal cues for depression analysis has been demonstrated in various depression detection/recognition tasks [7, 8, 72, 131–133]. Besides, visual cues like facial expression and facial dynamics have also proven to be effective in depression analysis [23,47,134,135]. This chapter investigates facial depression recognition, aiming to predict the depression level for a given face video based on the BDI-II metric [136].

While encouraging progress has been made over the past few years, automated depression analysis in videos remains challenging due to the following reasons:

On one hand, unlike those large-scale image datasets (e.g., ImageNet [137]) for visual recognition [45], the size of most existing depression datasets (e.g., AVEC 2014 [18]) is relatively small due to the privacy concerns. While representation learning based on CNN has been proved to be more effective than hand-crafted descriptors in visual-based depression recognition [47], the lack of labeled data makes the model training with deep networks prone to over-fitting in practice.

On the other hand, many learning methods in the literature have been devoted to multimodal fusion of audio and/or video features for depression recognition [47, 72, 135], which have demonstrated boosted recognition performance by exploiting the complementary information encoded in different modalities. However, essential correlation and diversity between different visual modalities have not been well investigated in previous visual-based methods, especially for multimodal fusion of visual cues with deep CNN architecture.

In this chapter, we propose a multimodal deep learning approach for facial depression recognition to address these issues. Specifically, a sequential fusion of facial appearance and dynamics is introduced to facilitate such multimodal representation learning. Here, facial appearance and dynamics are adopted as the basis modalities in our multimodal fusion framework, as they have been validated to be effective in visual-based depression diagnosis [47]. Hence, a fusion between the two modalities is first operated on the blocks of a two-stream CNN. By such fusion of mid-level features in the CNN training, an initial interaction is conducted

to optimize the complementary patterns. Then, the extracted feature, together with the predicted label from the first stream (e.g., RGB modality), is fed into the second stream (e.g., Optical Flow modality) to refine the final prediction. We show that such sequential fusion can provide a probabilistic perspective about model correlation and complementarity between two different data modalities for improved depression recognition. We conduct experiments on the benchmark dataset (AVEC2014), and the results show the superiority of our method against several state-of-the-art alternatives. The main contributions of this chapter are:

- We proposed a sequential chained-fusion approach for depression recognition. With a probabilistic perspective, the proposed approach models the correlation and complementarity between facial appearance and facial dynamics at several network layers, such that the complementary and correlation information of different visual cues extracted from videos could be well exploited in model learning (Section 5.2).
- We evaluate our approach on the benchmark dataset and empirically show improvement over several state-of-the-art alternatives (Section 5.3).

The rest of this chapter is organized as follows. As the related work is listed in 2, we first give the detail of the proposed sequential fusion method in Section 5.2. Experimental settings, results and discussions are presented in Section 5.3, and Section 5.4 concludes the chapter.

5.2 Our Approach

Predicting the severity of facial depression is a process of learning spatio-temporal features related to human emotion categorization [138] from face videos. The facial appearance of a subject is one of the important visual patterns for depression recognition. At the same time, facial dynamics characterized by optical flow captures the temporal variations of appearance across frames. As such, we propose a sequential fusion approach to investigate the correlation and complementarity between two different data modalities for depression analysis. As shown in Fig.

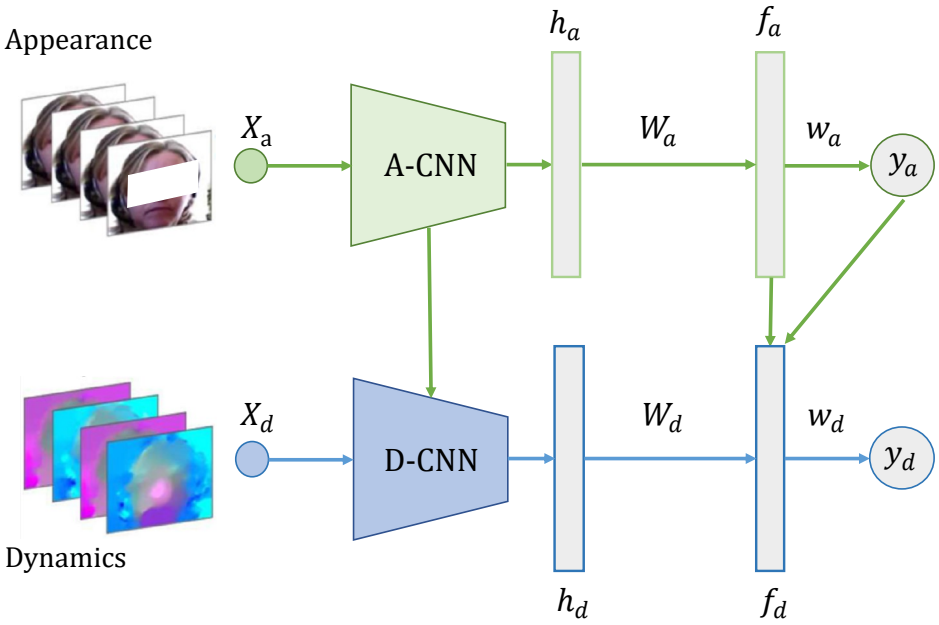


Fig. 5.1: Overview of our proposed sequential fusion approach for facial depression recognition. The (mid- and high-level) features extracted from the first stream (RGB modality) together with the predicted label are fed into the second stream (dynamics modality) for refinement of the final prediction.

5.1, we use a two-stream network architecture, where the encoders for each stream have the same backbone structure (can be any off-the-shelf CNNs).

Fundamentally, depression estimation can be viewed as a regression task, and hence we employ the mean square error (MSE) as the loss function. Mathematically, the loss L is defined as:

$$L = \frac{1}{2N} \sum_{i=1}^{N} \|y_i - \hat{y}_i\|^2, \tag{5.1}$$

where N is the number of the samples, $\hat{y}_i$ is the output prediction of the second stream of our network for the ith sample, and y_i is its ground-truth label.

5.2.1 Appearance-CNN

CNNs have been proved with powerful capability on image classification tasks over large-scale image data. Conversely, CNN is not a proper option to capture features from the dataset with a limited size. For depression estimation, available

datasets are usually with limited data and subjects. To handle this issue, we employ the pre-training and fine-tuning strategies to train the facial appearance CNN (A-CNN).

Due to time restrictions or computational restraints, it's not always possible to build a deep model from scratch which is the reason why we use pre-trained model. To achieve facial representations, we train two pre-trained deep networks (e.g., GoogLeNet [139] and ResNet-50 [38]) over CASIA-WebFace database [140], which is a public face recognition database containing 494,414 facial images from 10,575 subjects.

After the pre-training step, we can obtain the general facial features through the pre-trained model while those features are not relevant to facial depression. Hence, depression data are fed into the pre-trained model for fine-tuning, such that the final model is capable of accurate depression estimation.

5.2.2 Dynamics-CNN

Along with facial appearance, facial dynamics is also an indispensable component in our proposed model. The dynamics-CNN (D-CNN) is built upon the same backbone as the A-CNN with the optical flow data. Unlike the static RGB data, facial dynamics model the motion patterns inherent in faces that can be highly indicative for visual depression analysis.

We compute the optical flow with the duality-based approach [141], which is a decent method with sufficiently fast speed. To feed the optical flow data into the D-CNN, we transform the optical flow signal into a 3-dimensional image data (x, y, m), where x, y and m represent the x-component, y-component and the magnitude of the flow, respectively. For better observation and calculation, we multiply each image by 16 and convert it to the closest integer between 0 and 255 [142].

Similarly, we employ the same architecture, configurations, and loss used in the appearance CNN to train the dynamics CNN. The difference lies in that the input to the dynamics stream is the *flow* image computed from video frames, and there is no pre-training step in D-CNN.

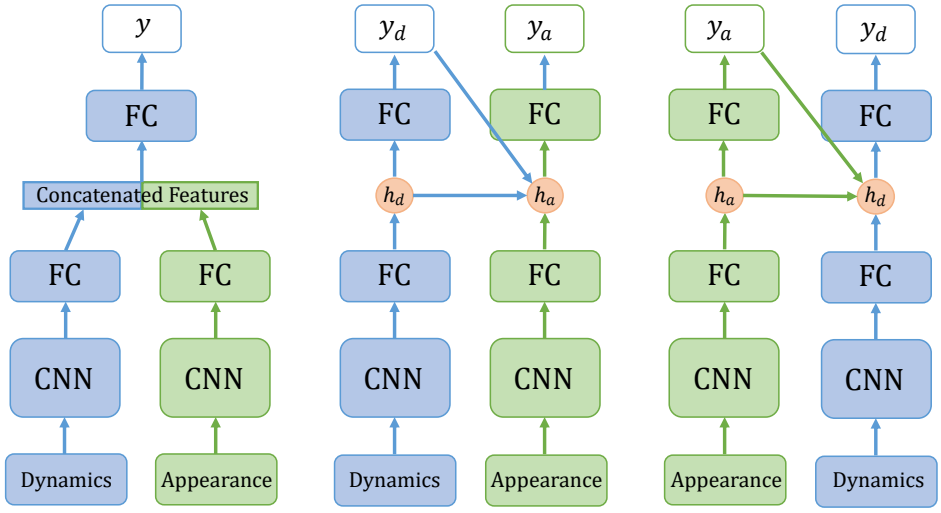
5.2.3 Sequential Fusion

Since Simonyan et al. [143] proposed a simple fusion framework with two separate convolutional networks on raw images and optical flow, respectively, multi-stream architectures entered the public consciousness and became a popular approach to handle multimedia tasks.

The baseline fusions are illustrated in Fig. 5.2, where the solution (a) applies the normal concatenation to form the input features of a set of fully-connected layers, which combines features from two streams directly, and the two CNNs are independent. In such a case, the learning manners of the two streams exert a slight influence on each other. Furthermore, as the input to the dynamics, CNN is the optical flow between several sequential frames; the changes between continuous frames are minor and difficult to distinguish. In order to achieve a compact and discriminative multimodal deep representation, a proper fusion method is needed to gather features from facial appearance and dynamics.

To integrate the two individual deep networks mentioned above, we introduce a sequential fusion approach based on a two-stream architecture. By making use of the features from both modalities, a Markov chain is established to integrate the two streams, which may refine the depression prediction sequentially. Considering different modality as the main input, a refined prediction is achieved by combining the hidden features (e.g., high-level features) and the predictions from the previous stream (see Fig. 5.2(b)-(c)). However, such fusion is made only after the FC layers, which means the *mid-level features* before the FC layers have not been exploited for better fusion. With the limited promotion of the feature fusion, the refinement of the predictions is improved, though the improvement is not significant.

Unlike the baseline fusion schemes shown in Fig. 5.2, our proposed sequential fusion (shown in Fig. 5.1) perform feature fusion not only after the FC layer, but also on the blocks of CNNs, by which both high-level and mid-level features of the CNNs can be well exploited to model the correlation and complementarity between different data modalities. In what follows, we give a probabilistic interpretation for such a fusion mechanism.



(a) Normal Concat (b) Appearance (main) + Dynamics (c) Appearance + Dynamics (main)

Fig. 5.2: Different fusion baselines for facial depression recognition. (a) Normally concatenated features from the dynamics and appearance streams; (b) The features extracted from the dynamics stream together with the predicted label are fed to the appearance stream for final prediction; (c) The features extracted from the appearance stream together with the predicted label are fed to the dynamics stream for final prediction.

For different input modalities, we assume that the depression predictions are conditionally independent. Consequently, we can factorize the joint probability into the conditional probabilities according to the conditional independence property. In a Markov chain, we predict the outputs $Y = \{y_1, y_2, \ldots, y_S\}$ on the given sequence of inputs $I = \{I_1, I_2, \ldots, I_S\}$ with $P(y|I)$ maximized. Due to the Markov property, we have

$$P(y|I) = P(y_1|I) \prod_{s=2}^{S} P(y_s|I, y_1, \ldots, y_{s-1}) \tag{5.2}$$

To model the likelihood in (5.2), the hidden feature and the prediction probability are respectively denoted by

$$\begin{aligned} h_s &= f([h_{s-1}, CNN(I_s), (y_1, y_2, \ldots, y_{s-1})]), \\ P(y_s|I, y_{<s}) &= \mathcal{N}(w_s^T h_s; \bar{y}_s, \sigma^2) \end{aligned} \tag{5.3}$$

where $s \in \{1, 2, \ldots, S\}$, f is an activation function, h_{s-1} denotes the hidden feature from the previous stream, y_s denotes the prediction of the sth stream, w_s denotes the regression coefficient vector, $\bar{y}_s$ is the ground-truth label, and σ is a certain standard deviation of the Gaussian $\mathcal{N}$. Here, $CNN(\cdot)$ denotes the convolutionary part and the first FC layer of the network, which can be any off-the-shelf backbones (e.g., VGG and ResNet).

In the proposed approach, the prediction of the dynamics stream is made conditioned on the appearance stream as well as the input dynamics data, which means that the final prediction is effected not only by the input of the current stream but also by the features and predictions from previous streams (see Fig.5.1). When the input modalities is $I = \{I_a, I_d\}$, we have:

$$\begin{aligned} h_a &= CNN(I_a), \\ P(y_a|I) &= \mathcal{N}(w_a^T h_a; \bar{y}_a, \sigma^2) \end{aligned} \tag{5.4}$$

and

$$\begin{aligned} h_d &= f([h_a, CNN(I_d), y_a]), \\ P(y_d|I, y_a) &= \mathcal{N}(w_d^T h_d; \bar{y}_d, \sigma^2) \end{aligned} \tag{5.5}$$

As known to all, maximization of $P(y_d|I, y_a)$ is equivalent to the minimization of the MSE loss defined in Eq. (5.1). Hence, our sequential fusion mechanism has a clear probabilistic interpretation.

It should be noticed that depressed patients are usually slow to initiate actions with stiff facial expressions. For depression analysis, motion features could be more discriminative than appearance cues. It is observed in our experiment that using the dynamics stream as the main stream to fuse the appearance stream can perform better prediction, and it also facilitates the training of the two-stream network. In the inference stage, only the prediction of the mainstream is used as the final depression prediction.

5.3 Experiments

To validate the effectiveness of our depression recognition approach, we conduct experiments on the AVEC 2014 benchmark dataset and compare its performance

with several state-of-the-art algorithms as well as the baselines. In what follows, a description of the dataset, data pre-processing, and experimental setting are first presented. Then, we present the results and analysis.

5.3.1 Dataset

We conduct the experiments on a database of the Audio/Visual Emotion Challenge (AVEC) 2014 [18], which is the most widely-used depression sub-challenge database for depression recognition. In AVEC dataset, the severity of depression is evaluated by the BDI scores which represent the depression level from 0 to 63, where the lower score represents more mild symptoms. The score evaluated in [0, 13], [14, 19], [20, 28] and [29, 63] indicate minimal, mild, moderate and severe depression, respectively. For each video clip, there are 3∼5 annotators predicting the BDI score.

In AVEC 2014, there are 84 subjects, and each subject needs to perform two different tasks named "Northwind" and "FreeForm" according to the instructions. All subjects in the two tasks speak German. There are 150 videos for each task, and the recordings were equally split into three partitions: training, development, and test set. Each partition includes 50 videos and has similar distributions in terms of gender, age, and depression levels for the partitions. All videos are recorded by webcam in a human-computer interaction scenario, and each video is approximately 2-minute length on average. There are at least three annotators per clip, and most clips are annotated by 5.

5.3.2 Data Pre-processing

As the raw data have a certain degree of noisy and redundant information which is irrelevant to depressive expressions. To extract meaningful information from noise, it is necessary to apply multiple pre-processing steps on the raw data before feeding it to the model. To avoid the waste of computing resources and speed up the training, subsampling is performed on the video frames with an interval of 10 frames which is determined experimentally.

To deal with the raw data, face detection and landmark localization of each subject in the video are implemented by Dlib [144]. After that, the facial region of an image size of 256×256 is cropped and aligned according to the eye locations.

After the above steps, we compute optical flow over a sequence of facial regions extracted from each video clip. As the input to dynamics stream, optical flow is computed between two frames which can capture facial motions known as face "flow images". A "flow image" has three components (x, y, m), where the first two channels are x, and y flow values and the third channel is the magnitude of the optical flow normalized between 0 and 255 with a median of 128.

5.3.3 Experimental Setting

We use two popular network architectures, i.e., VGG-11 [145] and ResNet-50 [38], as the backbone of our two-stream network to train the appearance and dynamic CNNs. In our experiments, the appearance CNN is pre-trained on a large-scale face dataset CASIA WebFace [140], while the dynamics CNN is trained from scratch. The MSE is adopted as the loss function for our depression regression.

The total number of training iterations for the appearance and dynamics CNNs are 400,000 and 600,000, respectively. We set the initial learning rate to 0.001, and decrease the learning rate by polynomial decay with power equals to 0.5, and set the momentum to 0.9 with a weight decay of 0.0002. For the joint training, the total number of iterations is 200,000 with an initial learning rate of 0.0001.

We use the mean absolute error (MAE) and root mean square error (RMSE) to measure the overall recognition performance. They are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{5.6}$$

and

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{5.7}$$

where N is the number of data samples, y_i and $\hat{y}_i$ are the ground truth and the prediction for the ith sample.

Table 5.1: Depression prediction results with different backbones on AVEC 2014

Model	MSE	RMSE
Appearance CNN(ResNet-50)	6.60	8.88
Dynamics CNN(ResNet-50)	8.64	10.71
Normal Concat(ResNet-50)	6.72	8.68
Appearance(main)+Dynamics(ResNet-50)	6.71	8.58
Appearance+Dynamics(main)(ResNet-50)	6.41	8.70
DeepFusion(ResNet-50)	**6.16**	**8.13**
Appearance CNN(VGG-11)	8.19	10.34
Dynamics CNN(VGG-11)	9.54	11.49
Normal Concat(VGG-11)	9.54	11.50
Appearance(main)+Dynamics(VGG-11)	8.41	10.89
Appearance+Dynamics(main)(VGG-11)	8.05	10.27
DeepFusion(VGG-11)	**7.54**	**9.79**

5.3.4 Experimental Results

5.3.4.1 Performance of Individual Models

We first investigate the impact of different fusion models (see Fig. 5.2) in our fusion framework for depression analysis. Six baselines are defined for evaluation of the performance of individual models: 1) Appearance CNN, 2) Dynamics CNN, 3) Normal Concat: fusion by normal concatenation on the FC layer of the two-stream CNN, 4) Appearance (main)+Dynamics: Dynamics stream is fed to Appearance stream for sequential fusion, 5) Dynamics (main)+Appearance: Appearance stream is fed to Dynamics stream for sequential fusion, 6) DeepFusion: Appearance stream is fed to Dynamics stream for a sequential fusion of both mid-level and high-level features.

As shown in Table 5.1, our proposed DeepFusion architecture built upon ResNet-50 achieves the MAE/RMSE of 6.16/8.13, which consistently outperforms the other baselines. Specifically, two sequential fusion models (dynamics CNN as the mainstream) perform better than other individuals, indicating the efficacy of such fusion mechanism. On the other hand, we can see that appearance CNN as the mainstream for sequential fusion may not be suitable for depression analysis, as the motion cues may play a more important role in features fusion for

final prediction. Finally, the proposed DeepFusion performs better than Dynamics (main)+Appearance, indicating that integration of mid-level features in the sequential fusion framework can further boost the overall prediction performance.

In addition, we investigate the impact of different backbones in our proposed approach. As shown in Table. 5.1, prediction models built upon ResNet-50 [38] consistently perform better than those built upon VGG-11 [145]. Also, our proposed DeepFusion achieves the best performance on both backbones in terms of MAE and RMSE.

5.3.4.2 Comparison with Previous Methods

We compared our proposed approach with several state-of-the-art depression recognition methods, and the results are presented in Table 5.2. For a fair comparison, the compared eight methods are based on the visual modality (e.g., face videos). Specifically, six of them are also deep learning-based solutions, and the other two are shallow learning models with hand-crafted video descriptors.

In [18], the baseline model for AVEC 2014 employed the epsilon-SVR with intersection kernel [146] trained using LGBP-TOP features. On the AVEC 2014 dataset, our approach beats the baseline approach by dropping the MAE by 2.70 and RMSE by 2.83. In [147], an SVR trained with the dynamic feature descriptors MRLBP-TOP and DPFV achieved the MAE/RMSE of 7.21/9.01. Also, our approach outperforms this shallow learning-based solution by a significant margin.

Our approach achieves the second-best performance on the AVEC 2014 dataset when compared to other seven deep learning-based works [79–81, 148–151]. In [148], the SlowFast networks transferred from action recognition model achieved the MAE/RMSE of 6.78/8.40. In [149], the combination of the global and local Convolutional 3D networks achieved the MAE/RMSE of 6.59/8.31. In [79], the MAE/RMSE of 6.86/8.78 was achieved by utilizing Bi-LSTM, whose input is the output of a deep CNN and TMP. In a very recent work [151], both local and global attention CNN are introduced for depression recognition and reduced the MAE/RMSE to 6.51/8.30. In [80], deep depression representation with visual explanation achieved the MAE/RMSE of 6.60/8.88, and later in [81], the deep

Table 5.2: Comparison with previous methods on AVEC 2014

Methods	MAE	RMSE
LGBP-TOP+SVR [18] (2014)	8.86	10.86
MRLBP-TOP+DPFV+SVR [147] (2018)	7.21	9.01
SlowFast Networks [148] (2019)	6.78	8.40
C3D(Global+Local) [149] (2019)	6.59	8.31
VLDN+Bi-LSTM+TMP [79] (2020)	6.86	8.78
DepressNet [80] (2020)	6.60	8.88
DJ-LDML [81] (2020)	6.59	8.30
Spectral-Representation [150] (2020)	**5.95**	**7.15**
DLGA-CNN [151] (2021)	6.51	8.30
DeepFusion (proposed)	6.16	8.13

metric learning-based solution achieved the MAE/RMSE of 6.59/8.30. Our approach consistently outperforms the aforementioned deep learning-based methods in terms of MAE and RMSE. The best-performed solution among the compared methods is the spectral representation of behavior primitives [150], which achieved the MAE/RMSE of 5.95/7.15. Different from our focus on fusion strategy for depression prediction, they used human behaviour primitives as the descriptor for each video frame and proposed spectral representations to represent video-level multi-scale temporal dynamics of expressive behavior. The superiority of their solution can be attributed to the consideration of the context of the measured behavior in depression representation learning.

5.4 Conclusion

In this chapter, we have proposed a deep multimodal learning method for the representation fusion of facial appearance and dynamics. To model the correlated and complementary depression patterns in multimodal learning, a chained-fusion mechanism is introduced to jointly learn facial appearance and dynamics in a unified framework. We have shown that such sequential fusion provides a clear probabilistic perspective of the model correlation and complementarity between two different data modalities for improved depression recognition. Experimental results on a benchmark dataset demonstrated the efficacy of our method when

compared to several state-of-the-art alternatives. In future work, investigation of the private-share model for multimodal depression representation learning appears to be an interesting topic.

Chapter 6

Conclusions and Future Directions

6.1 Conclusions

In this dissertation, I present a study on image sentiment analysis involving statistic image, dynamic image and text-image joint data on images and especially on human facial images.

For images from social media, we firstly explore the feasibility of learning visual concepts automatically via utilizing a deep CNN to capture visual features and a low-dimension RNN to deal with sequential concepts. In the deep CNN, CRBM is employed in each layer for visual feature extraction; the inter-connected hidden neurons of RNN are operated to delve into the potential relevance of sentiment concepts. For classic image analyzing techniques, our model with RNN is capable to deal with images in sequence format, Which means our model can be adapt for animated GIFs so far as to multimedia containing images. Compared with approaches training models using text-image (or audio-image) joint datasets, our approach shows a reasonable detection accuracy. Moreover, our approach shows superiority in training steps since we need image data only.

In our second work, we conduct the attempt to learn sentiment features from images with additional text information. A novel sentiment score is proposed by combining the image and text predictions. we identify the contents and sentiments in images through the fusion of both image and text features. Leveraged

on the fact that AlexNet is a pre-trained model with great performance in image classification and the corresponding set of images are extracted from the web, we present a novel method to extract features from Twitter images and the corresponding labels or tweets using deep CNN trained on Twitter data. By fine tuning pre-trained AlexNet, an initialized model is employed with text features extracted by AffectiveSpace in terms of English concepts. Lastly, to combine the image and text predictions we propose a novel sentiment score. Our model is evaluated on Twitter dataset of images and corresponding labels and tweets. We show that accuracy by merging scores from text and image models is higher than using any one system alone.

As a subdomain of visual sentiment analysis, the application of depression recognition showed the validity and significance in clinical diagnosing and monitoring depressive patients. we propose a sequential fusion method for facial depression recognition. For mining the correlated and complementary depression patterns in multimodal learning, a chained-fusion mechanism is introduced to jointly learn facial appearance and dynamics in a unified framework. We show that such sequential fusion can provide a probabilistic perspective of the model correlation and complementarity between two different data modalities for improved depression recognition. Results on a benchmark dataset show the superiority of our method against several state-of-the-art alternatives.

6.2 Future Directions

Despite the considerable progress of image sentimental analysis in recent years, to assist in practical application, more efforts should be made to collect additional data, investigate a range of methods, design and implement automated systems for practical applications. At present, several research challenges remain to be addressed:

- The availability and accessibility of databases. Due to limited amount of the training samples in existing depression datasets (for example, only 300

videos available in AVEC 2014), construction of a large-scale public depression dataset appears to be an emerging task in this research field. On the other hand, a multi-modal dataset with audio, text, and video depression cues will allow more effective training of depression analysis models.

- Multimodal depression representation learning. Multimodal learning seeks to build models that jointly leverage information from multiple modalities. Multimodal learning methods fuse different modalities at different stages with early, intermediate or late fusion. While existing multimodal learning approaches integrate the complementary information of multiple modalities from different views, basically they are weak in modeling robustness which is important for safety-critical medical diagnosis domains. Multimodal depression representation learning with data noise, label noise, and domain shift remain challenging.

- Trustworthy sentiment analysis. Quantifying the uncertainty of artificial intelligence models has received increasing attention over the past decades, especially in safety-critical domains such as medical diagnosis. Uncertainty estimation gives an effective way for trustworthy prediction. In general, the decisions made by AI models without uncertainty estimation can be untrustworthy, as they are prone to be inaccurate by data noises or limited training data. In the field of image sentiment analysis, as far as we know, few attempts have been made for trustworthy modeling of ISA, which remains challenging both for classification (positive or negative prediction, for example) and regression (depression prediction, for example).

References

[1] Q. Chen, I. Chaturvedi, S. Ji, and E. Cambria, "Sequential fusion of facial appearance and dynamics for depression recognition," *Pattern Recognition Letters*, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865521002397

[2] K. Jacksi and S. M. Abass, "Development history of the world wide web," *Int. J. Sci. Technol. Res*, vol. 8, no. 9, pp. 75–79, 2019.

[3] J. Deonna and F. Teroni, *The emotions: A philosophical introduction*. Routledge, 2012.

[4] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[5] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, 2010.

[6] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[7] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: a multimodal approach," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 11–20.

[8] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–7.

[9] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 2003, pp. 179–186.

[10] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 105–114.

[11] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[12] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "Effective attention modeling for aspect-level sentiment classification," in *Proceedings of the 27th international conference on computational linguistics*, 2018, pp. 1121–1131.

[13] F. Xing, D. H. Hoang, and D.-V. Vo, "High-frequency news sentiment and its application to forex market prediction," in *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*, 2020.

[14] K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, and A. Hussain, "Sentiment analysis of persian movie reviews using deep learning," *Entropy*, vol. 23, no. 5, p. 596, 2021.

[15] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications*, 2021.

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[17] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 3–10.

[18] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th international workshop on audio/visual emotion challenge*, 2014, pp. 3–10.

[19] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.

[20] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks." in *AAAI*, 2015, pp. 381–388.

[21] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 439–448.

[22] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[23] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1432–1441, 2015.

[24] W. H. Organization *et al.*, "Depression and other common mental disorders: global health estimates," World Health Organization, Tech. Rep., 2017.

[25] M. Avinash and E. Sivasankar, "A study of feature extraction techniques for sentiment analysis," in *Emerging Technologies in Data Mining and Information Security*. Springer, 2019, pp. 475–486.

[26] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[27] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognition Letters*, vol. 125, pp. 264–270, 2019.

[28] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd international conference on Knowledge capture*, 2003, pp. 70–77.

[29] Y. Hu, F. Wang, and S. Kambhampati, "Listening to the crowd: Automated analysis of events via aggregated twitter sentiment." in *IJCAI*. Citeseer, 2013, pp. 2640–2646.

[30] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *EMNLP 2002, Volume 10*. ACL, 2002, pp. 79–86.

[31] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment analysis in tourism: capitalizing on big data," *Journal of Travel Research*, vol. 58, no. 2, pp. 175–191, 2019.

[32] Y. Fu, J.-X. Hao, X. Li, and C. H. Hsu, "Predictive accuracy of sentiment analytics for tourism: A metalearning perspective on chinese travel news," *Journal of Travel Research*, vol. 58, no. 4, pp. 666–679, 2019.

[33] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, "Sentiment analysis on product reviews using machine learning techniques," in *Cognitive Informatics and Soft Computing*. Springer, 2019, pp. 639–647.

[34] A. Picasso, S. Merello, Y. Ma, L. Oneto, and E. Cambria, "Technical analysis and sentiment embeddings for market trend prediction," *Expert Systems with Applications*, 2019.

[35] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.

[36] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 83–92.

[37] J. Yang, D. She, M. Sun, M.-M. Cheng, P. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, 2018.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[40] I. Mozetič, M. Grčar, and J. Smailović, "Multilingual twitter sentiment classification: The role of human annotators," *PloS one*, vol. 11, no. 5, p. e0155036, 2016.

[41] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[42] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark." in *AAAI*, 2016, pp. 308–314.

[43] D. Borth, T. Chen, R. Ji, and S.-F. Chang, "Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content." ACM, 2013, pp. 459–460.

[44] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[46] L. Zheng, Y. Yang, and Q. Tian, "Sift meets cnn: A decade survey of instance retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1224–1244, 2017.

[47] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 578–584, 2017.

[48] R. Plutchik, "Plutchiks wheel of emotions," 1980.

[49] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," *Cognitive behavioural systems*, pp. 144–157, 2012.

[50] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 715–718.

[51] Y. Zhang, L. Shang, and X. Jia, "Sentiment analysis on microblogging by integrating text and image features," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2015, pp. 52–63.

[52] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2837–2841.

[53] T. Narihira, D. Borth, S. X. Yu, K. Ni, and T. Darrell, "Mapping images to sentiment adjective noun pairs with factorized neural nets," *arXiv preprint arXiv:1511.06838*, 2015.

[54] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: tional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[55] Q. You, J. Luo, H. Jin, and J. Yang, "Joint visual-textual sentiment analysis with deep neural networks," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1071–1074.

[56] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction," in *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*. ACM, 2015, pp. 57–62.

[57] L. Wu, M. Qi, M. Jian, and H. Zhang, "Visual sentiment analysis by combining global and local information," *Neural Processing Letters*, vol. 51, pp. 1–13, 06 2020.

[58] E. Sanchez, M. K. Tellamekala, M. Valstar, and G. Tzimiropoulos, "Affective processes: stochastic modelling of temporal context for emotion and facial expression recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9070–9080.

[59] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *California mental health research digest*, vol. 8, no. 4, pp. 151–158, 1970.

[60] P. Ekman, W. V. Freisen, and S. Ancoli, "Facial signs of emotional experience." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1125, 1980.

[61] W. V. Friesen and P. Ekman, "Emfacs-7: Emotional facial action coding system," *Unpublished manuscript, University of California at San Francisco*, vol. 2, 1983.

[62] M. D. van der Zwaag, J. H. Janssen, and J. H. Westerink, "Directing physiology and mood through music: Validation of an affective music player," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 57–68, 2013.

[63] P. Pal, A. N. Iyer, and R. E. Yantorno, "Emotion detection from infant facial expressions and cries," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 2. IEEE, 2006, pp. II–II.

[64] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition," in *Artifical intelligence for human computing*. Springer, 2007, pp. 91–112.

[65] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 4, pp. 966–979, 2012.

[66] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-visual emotion recognition using an emotion space concept," in *16th European Signal Processing Conference, Lausanne, Switzerland*, 2008.

[67] H. A. Vu, Y. Yamazaki, F. Dong, and K. Hirota, "Emotion recognition based on human gesture and speech information using rt middleware," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE, 2011, pp. 787–791.

[68] N. Banda and P. Robinson, "Noise analysis in audio-visual emotion recognition," in *International Conference on Multimodal Interaction, Alicante, Spain*. Citeseer, 2011, pp. 1–4.

[69] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[70] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007, pp. 401–408.

[71] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[72] H. Meng, D. Huang, H. Wang, H. Yang, M. Ai-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 21–30.

[73] H. Meng and N. Pears, "Descriptive temporal template features for visual motion recognition," *Pattern Recognition Letters*, vol. 30, no. 12, pp. 1049–1058, 2009.

[74] S. De Jong, "Simpls: an alternative approach to partial least squares regression," *Chemometrics and intelligent laboratory systems*, vol. 18, no. 3, pp. 251–263, 1993.

[75] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 356–361.

[76] H. Pérez Espinosa, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyez-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 49–55.

[77] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 73–80.

[78] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 19–26.

[79] M. A. Uddin, J. B. Joolee, and Y.-K. Lee, "Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm," *IEEE Transactions on Affective Computing*, 2020.

[80] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 542–552, 2020.

[81] X. Zhou, Z. Wei, M. Xu, S. Qu, and G. Guo, "Facial depression recognition by deep joint label distribution and metric learning," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[82] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognition Letters*, vol. 125, no. 264-270, 2019.

[83] P. Chhokra, A. Chowdhury, G. Goswami, M. Vatsa, and R. Singh, "Unconstrained kinect video face database," *Information Fusion*, vol. 44, pp. 113–125, 2018.

[84] A. Wang, J. Cai, J. Lu, and T.-J. Cham, "Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1125–1133.

[85] E. Cambria, N. Howard, J. Hsu, and A. Hussain, "Sentic blending: Scalable multimodal fusion for continuous interpretation of semantics and sentics," in *IEEE SSCI*, Singapore, 2013, pp. 108–117.

[86] A. Wang, J. Lu, J. Cai, T.-J. Cham, and G. Wang, "Large-margin multi-modal deep learning for rgb-d object recognition," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1887–1898, 2015.

[87] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 681–687.

[88] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[89] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2904–2913.

[90] N. Srivastava, R. Salakhutdinov *et al.*, "Multimodal learning with deep boltzmann machines." in *NIPS*, vol. 1. Citeseer, 2012, p. 2.

[91] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis-hybrid architectures," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 239–253, 2021.

[92] Y. Shang, Y. Pan, X. Jiang, Z. Shao, G. Guo, T. Liu, and H. Ding, "Lqgdnet: A local quaternion and global deep network for facial depression recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[93] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[94] S. Mai, H. Hu, J. Xu, and S. Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 320–334, 2022.

[95] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[96] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of LREC*, vol. 6, 2006, pp. 417–422.

[97] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis," in *Data Engineering Workshop at ICDEW 2008*. IEEE, 2008, pp. 507–512.

[98] B. Ohana and B. Tierney, "Sentiment classification of reviews using sentiwordnet," in *9th. IT & T Conference*, 2009, p. 13.

[99] C. Strapparava and A. Valitutti, "Wordnet affect: an affective extension of wordnet." in *LREC*, vol. 4, 2004, pp. 1083–1086.

[100] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008, pp. 1556–1560.

[101] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American society for information science and technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[102] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[103] T.-Y. Chang, Y. Liu, K. Gopalakrishnan, B. Hedayatnia, P. Zhou, and D. Hakkani-Tur, "Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks," *arXiv preprint arXiv:2105.05457*, 2021.

[104] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "Commonsenseqa: A question answering challenge targeting commonsense knowledge," *arXiv preprint arXiv:1811.00937*, 2018.

[105] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "Senticnet: A publicly available semantic resource for opinion mining," *Artificial Intelligence*, pp. 14–18, 2010.

[106] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open mind common sense: Knowledge acquisition from the general public," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2002, pp. 1223–1237.

[107] E. Cambria, D. Rajagopal, K. Kwok, and J. Sepulveda, "Gecka: game engine for commonsense knowledge acquisition," in *The Twenty-Eighth International Flairs Conference*, 2015.

[108] E. Cambria, J. Fu, F. Bisio, and S. Poria, "Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis." in *AAAI*, 2015, pp. 508–514.

[109] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.

[110] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 12, pp. 1357–1362, 1999.

[111] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*. IEEE, 2000, pp. 46–53.

[112] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[113] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[114] M. L. I. C. Dy, I. V. L. Espinosa, P. P. V. Go, C. M. M. Mendez, and J. W. Cu, "Multimodal emotion recognition using a spontaneous filipino emotion database," in *Human-Centric Computing (HumanCom), 2010 3rd International Conference on*. IEEE, 2010, pp. 1–5.

[115] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 18–31, 2012.

[116] J. T. Olin, L. S. Schneider, E. M. Eaton, M. F. Zemansky, and V. E. Pollock, "The geriatric depression scale and the beck depression inventory as screening instruments in an older adult outpatient population." *Psychological Assessment*, vol. 4, no. 2, p. 190, 1992.

[117] C. Qian, I. Chaturvedi, S. Poria, E. Cambria, and L. Malandri, "Learning visual concepts in images using temporal convolutional networks," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2018, pp. 1280–1284.

[118] T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5528–5531.

[119] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 45–63, 2014.

[120] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009, pp. 609–616.

[121] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 223–232.

[122] C. Qian, E. Ragusa, I. Chaturvedi, E. Cambria, and R. Zunino, "Text-image sentiment analysis," 2019.

[123] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 245–250.

[124] D. Balduzzi, "Randomized co-training: from cortical neurons to machine learning and back again," *arXiv preprint arXiv:1310.6536*, 2013.

[125] T. Sarlos, "Improved approximation algorithms for large matrices via random projections," in *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*. IEEE, 2006, pp. 143–152.

[126] D. Achlioptas, "Database-friendly random projections: Johnson-lindenstrauss with binary coins," *Journal of computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.

[127] Y. Lu, P. S. Dhillon, D. P. Foster, and L. H. Ungar, "Faster ridge regression via the subsampled randomized hadamard transform," 2013.

[128] J. A. Tropp, "Improved analysis of the subsampled randomized hadamard transform," *Advances in Adaptive Data Analysis*, vol. 3, no. 01n02, pp. 115–126, 2011.

[129] S. Ross and N. Heath, "A study of the frequency of self-mutilation in a community sample of adolescents," *Journal of youth and Adolescence*, vol. 31, no. 1, pp. 67–77, 2002.

[130] J.-P. Lépine and M. Briley, "The increasing burden of depression," *Neuropsychiatric disease and treatment*, vol. 7, no. Suppl 1, p. 3, 2011.

[131] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013, pp. 41–48.

[132] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE transactions on affective computing*, vol. 4, no. 2, pp. 142–150, 2012.

[133] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 35–42.

[134] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and vision computing*, vol. 32, no. 10, pp. 641–647, 2014.

[135] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear, "Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 478–490, 2016.

[136] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, "Comparison of beck depression inventories-ia and-ii in psychiatric outpatients," *Journal of personality assessment*, vol. 67, no. 3, pp. 588–597, 1996.

[137] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[138] Z. Wang, S. Ho, and E. Cambria, "A review of emotion sensing: Categorization models and algorithms," *Multimedia Tools and Applications*, vol. 79, pp. 35 553–35 582, 2020.

[139] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[140] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[141] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l 1 optical flow," in *Joint pattern recognition symposium*. Springer, 2007, pp. 214–223.

[142] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3164–3172.

[143] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, pp. 568–576, 2014.

[144] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[145] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[146] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.

[147] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1476–1486, 2018.

[148] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6202–6211.

[149] W. C. de Melo, E. Granger, and A. Hadid, "Combining global and local convolutional 3d networks for detecting depression from facial expressions," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[150] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, 2020.

[151] L. He, J. C.-W. Chan, and Z. Wang, "Automatic depression recognition using cnn with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, 2021.