國立政治大學理學院資訊科學系
社群網路與人智計算 國際研究生博士學位學程
College of Science
Department of Computer Science
National Chengchi University
Taiwan International Graduate Program in Social Networks and Human-Centered Computing

博士學位論文
Doctoral Dissertation

使用圖像和深度學習了解社交互動
Understanding Social Interaction Using Images and Deep Learning

博士班學生: 艾費瑪　撰
Student: Fatma Said Abousaleh Abdeo

指導教授： 曹昱　博士
余能豪　博士
Advisors: Yu Tsao, Ph.D.
Neng-Hao Yu, Ph.D.

中華民國 110 年 01 月
January, 2021

DOI:10.6814/NCCU202100261

This thesis is dedicated to the sake of Allah (SWT), who has been my constant source of inspiration, wisdom, health, knowledge and skills during the challenges of my whole life. He also gave me the strength to complete my research work successfully when I thought of giving up. I also would like to dedicate this work to my beloved parents, who continually provide me with their moral, spiritual, emotional, and financial support.

DOI:10.6814/NCCU202100261

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Furthermore, I am aware of and understand the University's policy on plagiarism and I certify that this dissertation is my own work, and, to the best of my knowledge and belief, does not breach copyright law, and has not been taken from other sources except where such work has been cited and acknowledged within the text.

Fatma Said Abousaleh

January 2021

DOI:10.6814/NCCU202100261

Acknowledgments

The completion of this doctoral dissertation would not have been possible without the support of several people. I would like to express my heartfelt gratitude to all who in one way or another contributed to the completion of this thesis. First and foremost, I would like to thank Allah (SWT) for granting me the blessing, patience, health and strength to undertake this research task and enabling me to its completion. Thank you so much, Allah, I will keep on trusting you for my future life.

I would like to express my deepest appreciation and thanks to my advisor Dr. Yu Tsao for offering me such a great opportunity to join his research lab and for taking the daunting responsibility to conduct this PhD research under his supervision. He provided me with extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. On the academic level, He provided all the necessary research facilities and the enlightening work environment that ultimately helped me to carry out and complete this study successfully. He also taught me how to research a problem and achieve goals. On a personal level, he inspired me by his hardworking and

passionate attitude that made me belong and fit into the amazing and diverse culture of Taiwan. I am also extremely grateful for his generous financial support during this research work.

I would also like to express my warmest thanks to my co-advisor Dr. Neng-Hao Yu for encouraging me in all stages of this work and for his vital support in crucial times. His encouragement and belief in me have greatly assisted me to carry on my research work, develop my academic identity and formulate the thesis of this research successfully. He was very generous in sharing his wealth of knowledge and experiences in research, academic life and beyond. These experiences will be my primary guide during my life and academic career.

I am extremely delighted to express my gratitude to Taiwan International Graduate Program (TIGP), Social Network and Human-Centered Computing (SNHCC) program, and Institute of Information Science (IIS) of Academia Sinica for the PhD fellowship that provided the financial support and the facilities which were needed to execute my research work. I am also grateful to the research fellows of the Institute of Information Science (IIS) and the Research Center for Information Technology Innovation (CITI) in Academia Sinica for their valuable guidance, consistent assistance, and useful critiques throughout this study. I am also deeply thankful to the National Chengchi University (NCCU) and all the faculty members of the Department of Computer Science for their support throughout my study period there.

DOI:10.6814/NCCU202100261

I am profusely thankful to Dr. Mark Liao for his constant guidance, empathy and motivation during this study. He has always made himself available to listen and clarify my doubts despite his busy schedules and I consider it as a great opportunity to learn from his research expertise. He gave useful advice that helped in addressing some of the shortcomings in this thesis. Your advice on both research as well as on my career has been priceless. Thank you, sir, for all your help and support. I also wish to thank Dr. Wen-Huang Cheng for the valuable and helpful suggestions and comments that he has provided me during the period of his supervision on this PhD research. His vital advice significantly contributed to the quality of this research.

Special thanks are due to my labmates in Biomedical Acoustic Signal Processing Laboratory (Bio-ASP), Multimedia Computing Laboratory (MCLab), and to my TIGP-SNHCC colleagues for their prompt help, constructive suggestions and productive discussions during this PhD journey and for the pleasurable moments that we spent together in Taiwan. Similarly, I am thankful to the assistant of TIGP-SNHCC, Ms. Chia-Chien, for her wonderful services and unconditional help all the time during my stay in Academia Sinica and Taipei.

Nobody has been more important to me in the pursuit of seeing this achievement come true than the members of my family. I owe a lot to my parents for their love, care, sacrifices, confidence, encouragement, and prayers at every stage of my personal and academic life. They set a great example for me about how to live, study, work, and handle mental stress in the dynamic environment. Thank

you both for giving me the strength to reach for the stars and chase my dreams. I am also grateful to my brother and sisters, who have provided me with endless support and incented me to strive towards my goal. I would also like to thank all my other family members and all my friends for their outstanding emotional support and well wishes during this whole journey.

DOI:10.6814/NCCU202100261

中文摘要

人們通常能自然無礙地和他人互動，而社群訊號（social signal）是有效溝通的自然產物。然而如何讓電腦能分析、了解社交互動，並正確展現人類社群訊號的過程，仍舊是社群訊號處理（social signal processing, SSP）領域最大的挑戰之一。社交互動可以透過面對面或網路兩種不同的渠道進行。在面對面的互動中，人們常透過可觀察的非語言行為線索（例如：手勢、臉部表情、聲音表達、肢體動作和人際距離等）來了解社群訊號和行為並與他人互動。基於臉部圖像辨識的社交互動研究近來受到學術界極大重視，這是因為臉部圖像蘊含多樣化的臉部特徵，可以用來傳達關於年齡、性別、情緒和健康狀況的資訊。這些訊息在描述個人特質和社交溝通中扮演了重要的角色，其中，年齡尤其是影響我們日常社交互動最基本的因素之一。因此，根據臉部影像自動估計年齡的研究成為人工智慧領域的一項重要目標。雖然近幾年有巨大進展，但由於臉部樣貌的多變性取決於基因特徵、生活型態、臉部表情以及年齡等因素，這個研究課題仍屬於未解的難題。另一方面，網路互動包含了用戶如何透過社交平台如Facebook、Twitter、Instagram或Flickr等與他人互動。大部分的社交網路允許用戶創造並分享內容，也可以藉由不同的形式（例如：觀看、按讚或留言）與其他用戶創造的內容互動，從而產生

DOI:10.6814/NCCU202100261

大量含有用戶興趣、觀點、日常生活和互動資訊的社交內容。爆炸性成長的社群媒體內容和線上互動的行為，造成少數社交內容得到大量關注、受歡迎，但絕大多數則受到忽視。在社群媒體上不同種類的內容中，圖像已經成為用戶溝通的重要媒介，也導致用戶獲得的觀看次數或社交知名度產生變動。上述現象吸引了電腦視覺和多媒體領域的研究人員的興趣，並探究特定圖像受歡迎的原因，以及如何自動預測其受歡迎程度。然而，因為用戶獨特的偏好及其在社群媒體上互動歷程等其他因素，社群媒體上圖像受歡迎的程度仍然難以衡量、預測和定義。為此，本論文提出了一個架構，用以理解現實和線上世界的社交互動，來解決這些挑戰。

首先，本論文探討根據臉部圖像自動估計年齡的問題。傳統估計臉部年齡的方法，透過直接分析臉部資訊（例如：鼻子、嘴巴、眼睛等）來從一個人的照片決定其年紀。然而即使對人類來說，一眼看出某人的年紀本質上仍是一項艱鉅的任務。為了處理這個問題，本論文由人類認知過程發想，提出了一個比較深度學習（comparative deep learning）的架構。藉由比較輸入圖像與選定的參考圖像（基準組），決定那組比較年輕或年長，從而以臉部圖像估算年齡。我們用區域卷積神經網路（region-convolutional neural network, R-CNN）從輸入圖像與參考樣本中擷取臉部特徵。然後，為了估計年齡差距，我們用能量函數（energy function）從全連接層（fully connected layer）獲取資訊，產生了一組代表比較關係（年輕或年長）的建議。最後，在模型的預測階段收集所有建議並依多數決來判斷人的年紀。我們在FG-

DOI:10.6814/NCCU202100261

NET、MORPH和IoG資料集上的實驗結果顯示，我們提出的架構超越目前最頂尖的方法，且進步的幅度分別是在FG-NET的13.24%（平均絕對誤差）、MORPH的23.20%（平均絕對誤差）以及IoG的4.74%（年齡分組分類精準度）。

其次，本論文研究社群媒體上圖片受歡迎度預測的問題。隨著社群網路如Flickr、Facebook的興起，用戶常藉由分享他們的生活照片來互動。雖然每分鐘上傳了數十億張圖像到網路，但只有少部分能有超過百萬次的觀看量，其他則完全被忽略。即使是相同用戶上傳的不同照片也不會有相同的觀看數。所以如何預測圖像受歡迎度是一個值得研究的主題，同時也是社群媒體分析的關鍵挑戰。因為這可提供一個瞭解個人喜好以及公眾目光的管道。然而，圖像受歡迎度的關鍵因素，和建立一個能預測社群媒體上圖像歡迎度的模型，依然是未解的難題。為此，本論文提出了一個多模式深度學習模型（multimodal deep learning），該模型藉由與圖像受歡迎度有關的多種視覺和社會特徵，來預測社群媒體上圖像的受歡迎度。本模型使用了兩種CNN，分別學習輸入圖像的高階特徵，並將他們融入一個統一的網路來預測受歡迎度。我們透過一系列對Flickr真實資料集的實驗來評估本模型的效能。實驗結果顯示，本預測模型勝過四個傳統的機器學習演算法、兩個CNN模型和其他最新的方法，效能至少提昇了2.33%（斯皮爾曼等級相關係數）、7.59%（平均絕對誤差）以及14.16%（均方誤差）以上。

DOI:10.6814/NCCU202100261

Abstract

Human beings generally have the capability to interact easily with each other without any obvious effort, and social signals are the natural result of this effective communication. The process of providing computers with an equivalent capability that enables them to analyze and understand social interactions, and then properly represent human social signals, remains one of the greatest scientific challenges in the field of social signal processing (SSP). Social interactions can take place in two different ways: face-to-face or cyber. In face-to-face interactions, people commonly use observable nonverbal behavioral cues (e.g., gestures, facial expressions, vocalizations, postures, interpersonal distance, etc.) to understand and interact with the social signals and behavior of others. The problem of recognizing social interactions from face images has recently received significant attention from the research community. This is because facial images have a variety of facial traits that can convey information about an individual's age, gender, emotions, and physical health. These types of information are known to play a key role both in the description of individuals and social communication. In particular, age is one of the most fundamental

DOI:10.6814/NCCU202100261

attributes that affect our daily social interactions. Automatic age estimation from face images has therefore become a significant task in numerous applications of artificial intelligence. Despite the huge advances in the automatic age estimation from face images in recent years, it remains a challenging problem. This is because of the large variations in facial appearance that result from a number of different factors, including genetic traits, lifestyle, facial expressions, and aging. On the other hand, cyber interactions are related to how users interact with each other through social media websites, such as Facebook, Twitter, Instagram, and Flickr. Most social networks allow users to create and share content and interact with other user-generated content in different forms (e.g., by viewing, liking, or commenting). This results in massive amounts of social content that provide information about users' interests, opinions, daily activities, and interactions. The explosive growth of social media content and the interactive online behaviors between users make only a limited number of social media content attracts a great deal of user attention and become popular, while the vast majority of content is completely ignored. Among the different types of content generated by users on social media, images have become important media for communication between users, resulting in variations in the number of views they receive or their social popularity. This phenomenon has attracted researchers from computer vision and multimedia domains to explore the reasons why certain photos are considered popular and how to predict their popularity automatically. However, it is still difficult to measure, predict, or even define image popularity on social

DOI:10.6814/NCCU202100261

media because it is based on a user's preferences and many other factors that could affect user's social interactions on social media websites and lead to the popularity of content. To this end, this dissertation proposes a framework for understanding social interaction in the real and online world to address these challenges.

First, this dissertation addresses the problem of automatic age estimation from facial images. The conventional methods for facial age estimation normally determine the age of a person directly from his/her facial image by analyzing some facial information (e.g., nose, mouth, eyes, etc.). This means only the input image is utilized to estimate the person's age. However, telling someone's precise age at a glance without any reference information is essentially a challenging task even for humans. To address this problem and inspired by human cognitive processes, this dissertation proposes a comparative deep learning framework that estimates the age from the facial image by comparing the input image with a set of selected reference images (labeled baseline samples) to determine whether the input face is younger or older than each of the baseline samples. A specific deep learning architecture, namely a region-convolutional neural network (R-CNN), is used to extract facial information from both the input image and the baseline samples. Then, an energy function is exploited to aggregate the extracted information from the fully connected layer in order to estimate age comparisons. This results in a set of hints where each hint represents a comparative relationship (younger or older). Finally, the estimation stage aggregates all the set of hints

DOI:10.6814/NCCU202100261

and then votes on the number of hints for each label in order to estimate the person's age. Therefore, the age of the input person could be estimated by taking the label that received the most votes. The experimental results on the FG-NET, MORPH, and IoG databases demonstrate that the proposed model outperforms compared to the state-of-the-art methods, with a relative improvement of 13.24% (on FG-NET), 23.20% (on MORPH) in terms of mean absolute error, and 4.74% (on IoG) in terms of age group classification accuracy.

Second, this dissertation addresses the problem of image popularity prediction on social media websites. With an increasing number of social networks such as Flickr and Facebook, users often interact with each other by sharing photos of their daily lives. Although billions of images are uploaded to the internet every minute, only a few of these images receive millions of views and become popular, while others are completely ignored. Even the different images posted by the same user receive a different number of views. This raises the problem of image popularity prediction, which has become a key challenge in social media analytics, as it offers opportunities to reveal individual preferences and public attention. However, the challenge remains to investigate crucial factors that influence image popularity, as well as modeling and predicting the evolution of image popularity on social media. To this end, this dissertation proposes a multimodal deep learning model that predicts the popularity of images on social media by using various types of visual and social features that are associated with image popularity. The proposed model uses two dedicated CNNs to learn high-

DOI:10.6814/NCCU202100261

level representations separately from the input features and then merges them into a unified network for popularity prediction. The performance of the model was evaluated by performing a series of experiments on a real-world dataset from Flickr. The evaluation results reveal that the proposed prediction model outperforms four traditional machine learning schemes, two CNN-based models, and other state-of-the-art methods, with a relative performance improvement of more than 2.33%, 7.59%, and 14.16% in terms of the Spearman rank correlation coefficient, mean absolute error, and mean squared error, respectively.

DOI:10.6814/NCCU202100261

Table of contents

DOI:10.6814/NCCU202100261

DOI:10.6814/NCCU202100261

DOI:10.6814/NCCU202100261

DOI:10.6814/NCCU202100261

List of figures

DOI:10.6814/NCCU202100261

DOI:10.6814/NCCU202100261

List of tables

DOI:10.6814/NCCU202100261

Chapter 1

Introduction

1.1 Background of the Study

Humans have the capability to express and understand social signals (SSs) that are created during social interactions, such as agreement, disagreement, conflict, empathy, politeness, hostility, and any other way of behaving towards others that cannot be expressed using just by words but by nonverbal behaviors. They also have the ability to manage them in order to get along well with others. This range of abilities is called social intelligence, which is an aspect of human intelligence and the most essential indicator of success in life. Therefore, understanding how people can easily interact with the world and with each other considers one of the greatest scientific problems in the field of social signal processing (SSP).

SSP is a new research field that aims at providing computers with social intelligence that enables them to adapt and work properly in social settings. In particular, this field focuses on

DOI:10.6814/NCCU202100261

how machines can participate in social interactions by automatically modeling, analyzing, and synthesizing many of the nonverbal behavior cues that people utilize to express socially relevant information or social signals [1]. Nonverbal communication plays an important role in our daily life, where humans utilize nonverbal behavioral cues (e.g., posture, interpersonal distance, facial expressions, gestures, etc.) that they can easily sense with their eyes and ears for recognizing human social signals, and understanding social behaviors of others and then interact with them accordingly. Thus, the fundamental idea of SSP is that these kinds of cues can be detected with microphones, cameras, and any other suitable sensor, and they can be used as machine detectable evidence for automatic analysis and understanding of social behavior. This implies that SSP will bring computing closer to human-centered approaches that effectively deal with psychological and behavioral responses natural to humans. This will have a major impact on various domains of computing technology, such as human-computer interaction technologies because, interfaces will become more adept to social interactions with users [2], multimedia content analysis techniques because, the content will be analyzed on the basis of human perception of reality around them [3], computer-mediated communication because, the transmission will include the social cues necessary for establishing a natural contact with others [4], and any other domain where computers must seamlessly integrate the lives of people.

All potential nonverbal behavioral cues occurring in social interactions have been grouped into five main categories by psychologists, and have been referred to as codes [5]. These five codes are gestures and postures, vocal behavior, space and environment, physical appearance, and face and eye behavior. Of these, the behavior face and eye is a critical code, as the face is

DOI:10.6814/NCCU202100261

our straight and naturally preeminent way of communicating and comprehending someone's affective state and intents based on the facial expression [6]. In addition, faces convey information about age, gender, health qualities, personality, attractiveness, and emotions of an individual, that are useful sources in social signal processing [7]. This implies that facial behavior plays a key role in shaping perceptions during social interactions [7–9]. For example, age is one of the most essential signals that can be derived from the human face and is considered an important factor in interpersonal communication and interaction in our social life, as the perception of the age of our interlocutors can help us to determine and respond to the way in which we interact with them. In addition to daily life, the capability to estimate age is helpful in more particular contexts such as police testament or the selling of products authorized only from a specific age. However, the capacity of the human for age estimation is usually not as strong as for estimating other facial traits. Therefore, developing automatic facial age estimation systems that are comparable or even superior to the human ability in age estimation has become an attractive and challenging subject of research in recent years. Consequently, the first problem that we address in this dissertation is how to develop an automatic facial age estimation system based on basic concepts of human use to estimate the age so that it can accurately predict a person's age from his/her facial image.

On the other hand, the impact of technology, particularly social media, on human life has caused enormous changes in human behavior. These changes include numerous areas of human interaction, influencing the way people communicate, interact, work, think, do business, act, and react. We can simply say that social media has extensively influenced every facet of human life. The most important change in persons' behavior after the emergence of

DOI:10.6814/NCCU202100261

online social networks is the way they interact, and its range. Thousands of millions of users daily create and share vast amounts of content on online social networks, and interact with each other irrespective of time and location. The content created by users on social media provides information about users as well as their living environments, allowing us to access a user's preferences, opinions, and interactions. Thus, analyzing this content provides an opportunity to comprehend human behavior and can also be employed to improve the user services provided by these networks.

Simultaneously, the existence of more connections on online social networks brings more attention and visibility to people, which is called popularity on social media. Popularity is measured by the number of fans, followers, friends, retweets, likes, or any other metrics used to calculate engagement, and this depends on the type of social network. The interactions and reactions of users to the posted content play a fundamental role in information diffusion and the popularity of content on social media [10, 11]. Once the content is posted on a social network, it attracts a different amount of user interactions based on its importance, subject matter, publisher's credibility, time of publishing, etc. [12, 13]. Meanwhile, some contents succeed in attracting more users' interactions and becoming popular [14]. The popularity of content is generally measured by different metrics, such as the number of views, shares, likes, comments, etc. Predicting the popularity of content on social media (which can be text, audio, video, or image) has become a significant research topic, as it provides an opportunity to understand how users interact with online content and how information propagates across social networks. Therefore, the second problem that we address in this dissertation is the analysis of the popularity of content on social media and, more specifically, visual content

DOI:10.6814/NCCU202100261

such as images, to first explore the factors that can affect the social interactions of users on social media websites and lead to content popularity. Second, we design an efficient prediction model that can accurately predict this popularity.

1.1.1 Social Interaction and Facial Age Estimation

Age is one of the most significant elements of face-to-face interaction because the age of our interlocutor strongly identifies the way we interact with him/her [15]. In almost all cultures, people interact differently with younger and older people. For example, some studies revealed that youths tend to talk more slowly and loudly to older people [16]. It is therefore normal in daily life to estimate the age of people in order to interact with them in a proper manner. Humans can observe aging-related traits on faces, which helps them to predict the age of other individuals only by looking towards their faces. However, researchers who have worked on the process of age estimation by humans conclude that humans are not so precise in age estimation [17]. The main explanation for this is that different people of the same age can have different facial appearances due to varying rates of facial aging [18, 19]. Therefore, several automatic facial age estimation methods have been developed to compensate for humans weakness in age estimation.

Although the automatic estimation of human age from face images has recently received significant attention, it remains a challenging problem. In particular, there are many reasons that make automatic age estimation a non-trivia task. First, the aging process is uncontrollable and is influenced not only by the genetic traits of a person but also by several external

DOI:10.6814/NCCU202100261

factors, such as dietary habits, environmental conditions, lifestyle, living location, and health status [20]. Second, the gender of a person can have a significant effect on age estimation. Recent studies on facial aging have shown that the aging process differs in some respects between males and females [21, 22], including the appearance of facial hair like beards, increased thickness, facial vascularity, hormonal effects, and possible variations in fat and bone absorption rates throughout the life cycle. For example, the development of deeper wrinkles around the perioral area is higher in women than in men, as their skin has fewer appendages compared to men [23]. Third, from a technical point of view, males and females may have different discriminatory facial aging features shown in images because of the different extents of using makeups, cosmetic surgeries, and accessories [19, 24]. For example, many photos of the female face can likely show younger appearances than they actually [25]. Therefore, the extraction of general discriminative features for age estimation while decreasing the negative impact of individual differences remains an open problem. Fourth, the load of obtaining large-scale databases that cover a sufficient age range with chronological face aging images makes it harder to perform the estimation tasks [26]. Although web image mining can assist in the data collection process [27], it is usually difficult or even impractical to compile a large database for a large number of subjects who can supply a sequence of personal photos at different ages. Finally, the age estimation process is often affected by certain imaging conditions of face images, such as the variation of head pose, blur, illumination, expression, and occlusion. It is also influenced by camouflage caused by beards, moustaches, glasses, and makeup in the face images.

DOI:10.6814/NCCU202100261

Automatic age estimation involves the automatic labeling a face image with the exact age (year) or the age group (year range) of the human face. Traditional computer vision methods for age estimation from face images rely on the extraction of certain handcrafted features that are carefully designed to represent the aging information and subsequently use these features to train a classification/regression machine learning model to predict the age of the face [28, 29, 18, 19, 24, 30]. These methods achieve relatively good results if they can effectively extract the most relevant features of aging. This means that the performances of these methods rely heavily on feature engineering, which is time-consuming, costs a lot of human effort, and requires expert knowledge. In addition, the model can be brittle in the case that the selected features are not appropriate for the age estimation task. On the other hand, a different approach is taken by deep learning-based methods compared with handcrafted methods. While using deep learning, the model is left to automatically extract and learn appropriate features related to age by feeding it with several samples of facial images during the training process. The features of the model are then iteratively (and automatically) tuned using the error between the initial output of the model and the desired (real) output. CNN is one of the most important deep learning models that has been successfully used in face analysis tasks, especially age estimation. This is because CNN can efficiently capture high-level complex age-related visual features from raw input face images without any handcrafting and has strong robust adaptability to the noise in the image, which indicates that the final estimation of the age will be more accurate. Deep learning schemes and CNNs have therefore been used in recent studies on age estimation and have demonstrated superior performance compared to other traditional methods [31–35]. However,

DOI:10.6814/NCCU202100261

their performances depend on the training efficiency of deep architecture and the appropriate choice of deep learning parameters, which are difficult to achieve, especially because of the ill-conditioning problem of the deep neural network [36].

Most of the existing methods for facial age estimation so far rely on the biometric features extracted from the input facial image to estimate the person's age. This indicates that only the input image is used to estimate the age of a person. However, telling someone's precise age at a glance without any reference information is difficult even for humans. In practice, humans commonly infer the age of a person by learning to create links between a known age and the corresponding facial cues of a person. They then take the learnt information as a reference to judge if an unseen face is younger or older than the reference. The accuracy of age estimation of an unseen face increases whenever the number of available references increases. Thus, according to human cognitive processes [37], a more robust way of estimating facial age is likely to be in a comparative manner, that is, learning from a number of comparative relationships (a given face is younger or older than another face of known age). As a result, there is an increasing demand for the development of an automatic facial age estimation algorithm that simulates the process of age perception by humans and outperforms the human ability in age estimation. Therefore, motivated by the human perception of age estimation and the strength of CNNs in extracting effective and discriminative aging features from face images, this dissertation proposes an automatic facial age estimation system known as comparative region convolutional neural network (CRCNN) that can efficiently estimate the age of a person by using the input face image information in addition to reference information

DOI:10.6814/NCCU202100261

obtained from some reference face images with known ages as well as considering the limitations of the abovementioned methods in age estimation.

1.1.2 Social Interaction across Social Media and Popularity Prediction

Social interaction is steeply increasing through online social network websites. The rapid development of social networks has offered a variety of features to facilitate socialization on the internet. Users on these platforms can interact with each other by creating and sharing various forms of content, such as texts, photos, audios, and videos. This has contributed to explosive growth in social media content and has intensified the online competition for users' attention because only a small amount of social content receives the most attention and becomes popular, while others are completely unnoticed. The popularity of social media content reflects people's interests and provides opportunities to comprehend how users interact with online content and how information is disseminated through social media websites, which have a profound impact on social economic and governmental activities. Thus, modeling and predicting the popularity of social media content has become a significant research subject in social media analytics, and an essential task for supporting the design and assessment of a wide range of systems, from targeted advertising to effective search and recommendation services.

Images are one of the main visual content posted by users on social networks and have become important media for communication among them. The explosive growth in the number of images posted on social media and interactive behaviors between web users results

DOI:10.6814/NCCU202100261

in a variation in the number of views that these images receive or their popularity on social networks. Thus, this interesting phenomenon has attracted the research community to explore the factors that make certain images more popular than others, and how to automatically predict their popularity on social media. However, predicting the popularity of images on social media is a nontrivial and challenging task. The main difficulty lies in the fact that image popularity can be affected by different factors and features, such as visual content, text content, aesthetic quality, user, and time, which are intertwined during the cascade process. In addition, it is nontrivial to build a regression model that can integrate and process the various features contributing to image popularity and accurately predict it.

Recent studies have designed different types of features that evidently influence the volume of image popularity [38–41]. However, most of these studies rely only on some useful features (e.g., visual content, social context, and post metadata) for image popularity prediction, and ignore interactions between other valuable features (e.g., time and aesthetic quality). For example, users prefer to browse social media sites during a specific time of the day, such as weekend leisure time, which means that images posted at that time are more likely to receive a large number of views and become popular. Similarly, an image with a high aesthetic quality usually attracts the user's attention and obtains a higher number of views. Thus, in addition to visual content, social context, and post metadata features, time and aesthetic features are also essential for accurately predicting image popularity. With regard to predictive models, existing works mostly use simple machine learning models to predict image popularity [38, 39, 42, 43]. Although these models have achieved satisfactory prediction accuracy, they are not sufficiently powerful to capture and extract high-level representations

DOI:10.6814/NCCU202100261

from the various types of raw features associated with image popularity, which consequently affects the predictive accuracy of these models. In addition, these models require both time and skill to fine-tune their hyperparameters. This implies that developing a predictive model that can effectively handle multimodal information contributing to image popularity and accurately predict it is highly desired. Therefore, this dissertation proposes a multimodal deep learning system, called visual-social convolutional neural network (VSCNN), to address image popularity prediction on social media in an efficient way, that is, the proposed VSCNN system learn effective and high-level representations from various visual and social features that significantly influence image popularity and precisely predict it.

1.2 Motivation

In the real world, age estimation is a skill that we use in everyday life, and it also has an important influence on our daily social interactions. Several automatic age estimation systems are designed to estimate a person's age from his/her face image, as the estimation of age by humans is not as easy as for determining other facial information (e.g., gender, identity, or expression). Although these systems have achieved promising results, the problem of age estimation is far from being solved. The major difficulty lies in how to design aging features that remain discriminative despite the significant variations in facial image appearance. This implies that addressing the automatic estimation of human age from face images is a well-established and challenging problem. In addition, the automatic human age estimation using

DOI:10.6814/NCCU202100261

facial image analysis has numerous potential real-world applications. These applications include:

(i) Security system access control: With an increasing number of crimes and terrorist threats, security control systems have become increasingly important in our daily lives. With the help of a monitoring camera, an automatic age estimation system can be used in the surveillance of bars as well as alcohol and cigarette vending machines to stop under-aged people from entering bars or wine stores and to prevent them from purchasing alcoholic drinks or cigarettes [44]. Age estimation can also be used to deny children access websites with unsuitable materials or restricted movies [45, 18]. In addition, age estimation can also play an important role in controlling money transfer fraud from ATMs by monitoring a specific age group that the police have found to be more prone to fraud [46].

(ii) Age-specific human-computer interaction: Individuals belonging to different age categories have various criteria and demands related to the way they interact with computers. If an automated age estimator is used to determine the age of a computer user, both the computing environment and the user interface could be adjusted automatically to meet the needs of his/her age group [47, 48]. For example, interfaces based on colorful icons with appropriate illustrations can be activated when dealing with young kids, while interfaces based on icons with titles written in large fonts can be activated for older users.

(iii) Development of automatic age progression systems: Automatic age progression systems have the ability to simulate aging effects on new face images to predict how the person might look like in the future, or how he/she looked like in the past. Because automatic

DOI:10.6814/NCCU202100261

facial age estimation systems depend on their ability to comprehend and categorize changes in facial appearance because of aging, the methodology needed for this task could form the basis for designing automatic age progression systems [29]. In addition, age progression algorithms often require information related to the current age of an individual, and this emphasizes the essential role of facial age estimation systems in the development of automatic age progression systems.

(iv) Electronic customer relationship management (ECRM): ECRM uses modern internet-based technologies, such as chat rooms, blogs, emails, forums, and web sites, to efficiently manage the distinguished relationships with customers and communicate with them individually [46, 49]. As customers come from different age groups, they may have varied consumption patterns, preferences, and expectations for the products. Accordingly, automatic age estimation can be used by companies for monitoring market tendencies and customizing their products and services to satisfy the needs and desires of clients in various age groups. The issue here is how substantive personal information from all customers' age groups can be obtained and analyzed without infringing their privacy rights. However, a camera capturing pictures of the clients can collect demographic data by snapping the face images of clients and automatically estimating their age groups using an automated age estimation system. All of these can be done without violating the privacy of anyone.

(v) Biometrics: Age estimation is a kind of soft biometrics that provides additional information about users' identity [50, 51]. It can be utilized to supplement the main biometric features, such as the face, fingerprints, iris, voice, and hand geometry, to enhance the performance and effectiveness of a hard (primary) biometrics system. For example, the

DOI:10.6814/NCCU202100261

system in real face recognition or identification applications often needs to recognize or identify faces after a gap that has lasted for many years (e.g., passport renewal or border security), that highlights the importance of age synthesis [52–54]. With the help of a dynamic aging model, the facial recognition system can dynamically fine-tune its parameters by taking into consideration the differences in face structure or skin texture during the aging process. As a result, the efficiency of the system in the time gap could be substantially improved [55].

All the aforementioned recent application areas of automated age estimation imply the need for developing more precise age estimation systems.

In the online world, social media websites have been designed with the aim of facilitating and increasing social interactions among people on the internet. We can simply say that social media websites have altered the way we live and interact with. In addition, social media platforms have democratized the process of creating web content, allowing mere users to become creators and distributors of content. However, this has also led to massive growth in social content and has intensified the online competition for users' attention. This is because the interactive behavior of Web users often makes some of the content published on social media more popular than others. Therefore, there is a growing research interest in modeling and predicting the popularity of social media content [56, 57]. Predicting the popularity of social media content, especially visual content such as images, can help us understand public interest and attention behind user interactions. It can also facilitate several practical applications, such as online advertising [58, 59], online marketing, network dimensioning (e.g., caching and replication) [56], content retrieval [60], and politics. For example, in the case of online advertising, advertisers would like to be able to predict the number of

DOI:10.6814/NCCU202100261

views that a specific advertisement might produce on a particular website. Thus, if the popularity count is directly related to advertisement profits (such as with advertisements shown with YouTube videos), profits may be fairly precisely estimated ahead of time if all parties know how many views the video is expected to receive. In addition, in online marketing, the popularity prediction of a given product on a marketing company website provides a great opportunity for the company to make more strategic decisions, such as better managing their resources and more effectively targeting their ads. In general, both the customer benefits from a more pleasurable experience and the company benefits from a monetary saving or gain. However, popularity prediction is not an easy task because of the difficulty in modeling and exploring the various factors that contribute to the popularity of social media content and, more specifically, image popularity. In addition, the popularity of different social media content co-evolves over time, and this evolution may be described by complex online interactions and information cascades that are difficult to predict at the microscopic level [61–63]. This implies that developing a popularity prediction system that can accurately predict the popularity of social media content is challenging and an active area of research.

1.3 Contribution

In this dissertation, we explore and automatically estimate one of the factors that influence social interactions in the real world, which is the face age, as it considers one of the most important factors in interpersonal communication and interaction in our social life. Further-

DOI:10.6814/NCCU202100261

more, it is known that human behavior, preferences, and interactions are different at different ages, which indicates vast potential applications of automatic age estimation. Simultaneously, we study the social interactions of users to online content by exploring the factors that can affect the social interactions of users on social media websites and, more specifically, on Flickr site and lead to content popularity. Then, we design an efficient prediction model that can accurately predict this popularity. Such predictions can help improve user experience and service effectiveness. The main contributions of this dissertation are as follows:

- Motivated by the human cognitive process and the strength of CNNs in extracting effective and discriminative aging features from face images, we propose a novel comparative deep learning framework for facial age estimation, called comparative region convolutional neural network (CRCNN). In the proposed CRCNN framework, not only the input face image is used, but also several other reference face images of known age are taken as baseline samples to compare with the input face image. The advantage of this comparative approach is that, in addition to the input face image information, some other side information obtained from the baseline samples can be exploited to boost the estimation task, leading to a more accurate estimation. In addition, instead of using classical deep learning models, the region-convolutional neural network (R-CNN) is exploited to account for the spatial context of facial regions. Moreover, the method of auxiliary coordinates (MAC) is incorporated in the training process of our framework to reduce the ill-conditioning problem of the deep network and provide efficient optimization. The experimental results on the FG-NET, MORPH, and IoG databases demonstrate that the proposed model achieves a significant

DOI:10.6814/NCCU202100261

outperformance compared to the state-of-the-art methods, with a relative improvement of 13.24% (on FG-NET), 23.20% (on MORPH) in terms of mean absolute error, and 4.74% (on IoG) in terms of age group classification accuracy.

- Motivated by multimodal learning approaches, that uses information from various modalities, and the current success of convolutional neural networks (CNNs) in processing data from different modalities, we propose a multimodal deep learning framework for image popularity prediction on social media, called visual-social convolutional neural network (VSCNN). The proposed VSCNN framework uses dedicated CNNs for separately learning high-level representations from different types of features associated with image popularity, including multi-level visual, deep learning, social context, and time features, and then fused them using a merged layer into a unified network for further processing and obtaining the final prediction. The fusion process in our model becomes easy to execute and does not suffer from the data representation problem because the semantic vectors resulting from the dedicated CNN models usually have the same form of data. Furthermore, the robust interpretation of incomplete and inconsistent multimodal input becomes more reliable at later stages because more semantic knowledge becomes available from various sources, boosting the prediction process. The simulation results demonstrate that the proposed VSCNN model significantly outperforms state-of-the-art models, with a relative improvement of more than 2.33%, 7.59%, and 14.16% in terms of the Spearman rank correlation coefficient, mean absolute error, and mean squared error, respectively.

DOI:10.6814/NCCU202100261

1.4 Dissertation Organization

Chapter 1 provides background information and a general introduction to social interaction in the real world and on social media websites. It also defines the research problem and sub-problems addressed in this dissertation. Chapter 2 provides an extensive literature review regarding facial age estimation systems. It also offers a comprehensive literature review related to image popularity prediction techniques. In Chapter 3, we propose and develop a comparative deep learning framework that accurately estimates the facial age based on human cognitive processes. In Chapter 4, we propose and design a multimodal deep learning framework that predicts the image popularity on social media in an efficient way by combining several multimodal features that significantly influence image popularity. Chapter 5 summarizes this dissertation, highlights its research contribution, and provides an insight for future work.

DOI:10.6814/NCCU202100261

Chapter 2

Literature Review

2.1 Introduction

First, in this chapter, we review related work on automatic facial age estimation. Specifically, this chapter reviews the related literature on several descriptors used to extract and represent aging-related features from facial images. It also reviews the state-of-the-art techniques developed for age estimation from face images. Second, in this chapter, we review related work on image popularity prediction on social media. Specifically, this chapter reviews the related literature on many types of features that significantly affect image popularity. In addition, this chapter reviews the state-of-the-art prediction models designed to predict image popularity.

DOI:10.6814/NCCU202100261

2.2 Human Age Estimation from Face Images

The current age estimation systems utilizing face images usually comprise of two concatenated stages: aging-related facial feature extraction and age estimation techniques. Thus, we review the literature on age estimation according to these two stages.

2.2.1 Aging-Related Facial Feature Extraction

Most previous studies for facial age estimation focused on the extraction and fusion of different types of facial features. For example, Choi *et al.* [64] compared the performances of various methods (e.g., sobel filter, difference image between original and smoothed image, ideal high pass filter (IHPF), Gaussian high pass filter (GHPF), Haar and Daubechies discrete wavelet transform (DWT)) for extracting local features that can be used for detailed age estimation. Both [65] and [66] combined global and local features (e.g., active appearance models (AAM), Gabor filters, local binary patterns (LBP), Gabor wavelets (GW), and local phase quantization (LPQ)) to form hybrid features in order to have a better facial aging representation. Furthermore, Huerta *et al.* [67] used a fusion of textural and local appearance-based descriptors to achieve faster and more accurate results. Guo *et al.* [68] proposed the use of canonical correlation analysis (CCA) for jointly estimating the age with other facial information such as gender.

Meanwhile, other studies concentrate on extracting new features specially designed to estimate age [69, 70]. For example, Guo *et al.* [69] proposed using the biologically inspired features (BIF) for estimating human age from faces, that are generated based on a pyramid

DOI:10.6814/NCCU202100261

of Gabor filters. Geng *et al.* [70] introduced an approach named as AGing pattErn Subspace (AGES) for age estimation. In this AGES approach, they model the aging process using an aging pattern which is defined as a sequence of face images of the same person at different ages and sorted in the time order. For encoding the face images, the AAM [71, 72] is used to extract the feature vectors represent the face images in an aging pattern, indicating that the extracted feature combines both the shape and the intensity of the face images. In [73], aging face was represented by integrating AAM, LBP, and Gabor features, that are extracted from the face image. Furthermore, Suo *et al.* [74] proposed to design four graphical facial features, that is, topology, geometry, photometry, and configuration, based on their recent developed multi-resolution hierarchical face model [75]. Guided by this hierarchical model, instead of densely pursuing all filters over the image lattice, they applied particular filters to various parameters at different levels to extract these four types of features for age estimation. The authors in [76] incorporated the features of LBP histogram with main components of BIF, shape and texture features of AAM, and the projection of the original image pixels to principal component analysis (PCA) subspace, for representing the aging of the face image. Both [77] and [67] used the histogram of oriented gradients (HOG) [78] to represent facial features. Recent studies also proposed to use high-level complex age-related visual features extracted using deep learning techniques such as CNN for automatic age estimation [32, 79, 80]. These studies demonstrated that high-level semantic features designed based on deep neural networks architectures usually perform better than hand-crafted features.

DOI:10.6814/NCCU202100261

2.2.2 Age Estimation Techniques

After extracting and representing aging features, the subsequent step is to estimate the age. Age estimation can be considered as a particular task of pattern recognition. It can be approached as a multi-class classification problem when each age label is viewed as a class [45, 81, 82]. On the other hand, age estimation can be considered as a regression problem when age labels are viewed as sequential chronological series [24, 83, 84]. Thus, age estimation techniques have been divided into the following two categories: a) classification-based methods; and b) regression-based methods.

Regarding the classification-based methods, the existing studies have introduced several types of classifications models. For instance, Gao and Ai [48] used a fuzzy version of a classifier known as linear discriminant analysis (LDA) [85] to classify the face image as one of the following four coarse categories: baby, child, adult, and old. Lanitis *et al.* [45] presented a quantitative evaluation of the performance of different classifiers for automatic age estimation, including the nearest neighbor classifier, the artificial neural networks (ANN), and a quadratic function classifier. Ueki *et al.* [81] formulated the age estimation as an 11-class classification problem for the Waseda human-computer Interaction Technology-DataBase (WIT-DB) which has 11-class age-groups registered. They first built eleven Gaussian models from each 11 age-group in a low-dimensional 2DLDA+LDA feature space using the expectation-maximization (EM) algorithm. The age-group classification is then determined by fitting the test image to each Gaussian model and comparing the likelihoods.

DOI:10.6814/NCCU202100261

Han and Jain [86] used three different support vector machines (SVMs) to predict the age group (or exact age), gender, and race of a subject.

By considering age estimation as a regression problem, Lanitis *et al.* [29] examined the following three formulations for the aging function: linear, quadratic, and cubic, using 50 raw model parameters. A genetic algorithm is employed to learn the optimal model parameters from training face images of various ages. Guo *et al.* [19, 18] applied support vector regression (SVR) technique on age manifold learned with the orthogonal locality preserving projections (OLPP) method for age estimation. To fit aging manifold learned with the conformal embedding analysis (CEA) method, Fu *et al.* [83, 24] used a multiple linear regression function [87], which attains considerable improvements over some existing methods. A semidefinite programming (SDP) formulation is used by Yan *et al.* [88] to solve the regression problem for age estimation, in which the regressor is learned from uncertain nonnegative labels. The authors demonstrated that using SDP formulation for age regression provides much better results than the quadratic regression function and the multilayer perceptrons. However, the SDP is computationally very expensive particularly when the training set is large.

Recently, several deep learning-based techniques have been used for facial age estimation. For example, Takimoto *et al.* [89] integrated a multilayered neural network with the adapted retinal sampling mechanism in order to estimate facial age. Geng *et al.* [90] proposed a constructive probabilistic neural network for facial age estimation based on learning from label distributions. The CNNs have been used in different recent studies on age estimation as well [91–94]. Niu *et al.* [33] used ordinal regression and multiple-output CNN

DOI:10.6814/NCCU202100261

for age estimation. A series of sub-problems were transformed into binary classification from the ordinal regression then solved by CNNs as each output layer matching one sub-problem. Chen *et al.* [95] proposed a cascaded classification- regression framework for estimating the apparent age from unconstrained face images using deep convolutional neural networks (DCNNs). An error-correcting mechanism is also used to correct any erroneous age prediction.

In summary, all the previous works followed the conventional paradigm for facial age estimation, i.e., learning direct mappings between the extracted facial features and the associated age labels. These observations motivated us to develop our comparative approach using the deep learning method for estimating facial age.

Motivated by the human cognitive processes [37], it is arguable that a more robust approach to estimate a facial age is to be in a comparative manner, that is, learning from a number of comparative relations (a given face is younger or older than another face of known age). The development of our approach was also inspired by other ranking-based methods, such as Ranking SVM [96], RankBoost [97], and RankNet [98]. Ranking SVM [96] formulates learning to rank as the problem of classifying instance pairs into two categories: correctly ranked and incorrectly ranked. Experimental results of this method demonstrated that the algorithm performs well in practice, successfully adapting the retrieval function of a meta-search engine to the preferences of a group of users. Nevertheless, the losses (penalties) of incorrect ranking between higher ranks and lower ranks, and incorrect ranking among lower ranks are specified the same. This remark will cause problems for facial age estimation, as the youngest and oldest persons have entirely different facial information.

DOI:10.6814/NCCU202100261

RankBoost [97] is another ranking algorithm that is trained on pairs; it is similar to our work because it attempts to directly solve the preference learning problem rather than solve an ordinal regression problem. The results are provided using decision stumps as weak learners. RankNet algorithm [98] is easy to train and performs well on a real-world ranking problem with large amounts of data. In addition, RankNet explores the use of a neural network formulation. A probabilistic cost for training systems is also proposed to learn ranking functions using pairs of training examples. In this study, a novel ranking approach is presented through the proposed comparative framework for facial age estimation. First, a set of selected references, i.e., baseline samples, is introduced into the framework to make each rank more robust. Second, the proposed age estimation model is generated using the deep learning technique, providing effective features to rank each age based on the facial information. Finally, the younger/older comparison will help provide robust ranking by leaning similar facial information to estimate similar ranks; thus, the ranking will be better structured.

2.3 Image Popularity Prediction on Social Media

In recent years, predicting the popularity of social media content has received substantial attention [99–102]. Regarding image popularity prediction, the related studies differ in terms of the definition of the popularity metric (e.g., view, reshare, and comment counts); however, they all share the same basic pipeline consisting of extracting and testing several types of features that influence popularity, followed by applying a classification or regression model

DOI:10.6814/NCCU202100261

for prediction. Therefore, we review these studies by categorizing them according to the features and prediction models used.

2.3.1 Features Influencing the Image Popularity

The existing studies have primarily focused on investigating the relative effectiveness of various feature types for predicting image popularity, including social context, visual content, aesthetic, and time. For instance, Khosla *et al.* [38] demonstrated that image content (e.g., gist, color histogram, texture, color patches, gradient, and deep learning features) and social cues (e.g., number of followers or number of posted images) have a significant effect on image popularity. Gelli *et al.* [39] employed visual sentiment features along with context and user features to predict a succinct popularity score of social media images. They demonstrated that sentiment features are correlated with popularity and have considerable predictive power if they are used together with context features. Cappallo *et al.* [40] demonstrated that latent image features can be used to predict image popularity. They explored the visual cues that determine popularity by identifying themes from both popular and unpopular images. McParlane *et al.* [41] performed image classification using a combination of four broad feature types, that is, image content, image context, user context, and tags, to predict whether an image will obtain a high or low number of views and comments in the future.

Compared to the aforementioned approaches, relatively few studies have been conducted to demonstrate the effect of time and aesthetic features on image popularity. For instance, Wu *et al.* [103] developed a new framework called multi-scale temporal decomposition to predict

DOI:10.6814/NCCU202100261

image popularity based on popularity matrix factorization. They explored the mechanism of dynamic popularity by factoring popularity into two contextual associations, i.e., user-item context and time-sensitive context. Furthermore, Almgren *et al.* [104] employed social context, image semantics, and early popularity features to predict the future popularity of an image. Specifically, they considered the popularity changes over time by collecting information regarding the image within an hour of uploading and keeping track of its popularity for a month. Totti *et al.* [42] analyzed the effect of visual content on image popularity and its propagation on online social networks. Along with social features, they proposed using aesthetic properties and semantic content to predict the popularity of images on Pinterest.

We observe that most of the aforementioned studies rely only on a part of the useful features for image popularity prediction, and do not consider the interactions between other pertinent types of features.

2.3.2 Prediction Models

Regarding the models used for image popularity prediction, previous studies have introduced several types of machine learning schemes. Both [38] and [39] considered image popularity prediction as a regression problem in which support vector regression (SVR) [105] was used to predict the number of views that an image received on Flickr. Totti *et al.* [42] reduced the problem to a binary classification task and utilized a random forest classifier [106] to predict whether an image would be extremely popular or unpopular based on the number of reshares

DOI:10.6814/NCCU202100261

on Pinterest. Moreover, the authors in [43] predicted the number of views for an image on Flickr using a gradient boosting regression tree [107]. Although most of these prediction models perform satisfactorily, they tend to generate smoothed results, making the popularity of images with overly high or low scores difficult to predict accurately. In addition, it may be time-consuming to fine-tune the hyperparameters that significantly influence the performance of these models.

Recently, deep learning techniques have gained widespread attention and achieved outstanding performances in various tasks [108–111], owing to the capability of deep neural networks to learn complex representations from data at each layer, where they imitate learning in the human brain by abstraction. Nevertheless, insignificant effort has been expended for predicting image popularity using these techniques. In this regard, Wu *et al.* [112] proposed a new deep learning framework to investigate the sequential prediction of image popularity by integrating temporal context and attention at different time scales. Moreover, Meghawat *et al.* [113] developed an approach that integrates multiple multimodal information into a CNN model for predicting the popularity of images on Flickr. Although these studies have achieved satisfactory performances, they are not sufficiently powerful to capture and model the characteristics of image popularity. For instance, the authors in [113] investigated the effect of the visual content of an image on its popularity by utilizing only one feature obtained by the pre-trained InceptionResNetV2 model, whereas they ignored other important visual cues, such as low-level computer vision, aesthetics, and semantic features. Moreover, although it has been demonstrated that time features have a crucial effect on image popularity [103, 112], they were not considered in the proposed model. They also adopted an early

DOI:10.6814/NCCU202100261

fusion scheme for processing the proposed multimodal features, despite several studies having demonstrated that this scheme is outperformed by the late fusion scheme in processing heterogeneous information [38, 114].

To address the above issues, we analyze a real-world dataset collected from Flickr to identify and extract different kinds of features that are correlated with image popularity, including multi-level visual, deep learning, social context information, and time features. Based on the extracted features and motivated by the recent success of convolutional neural networks (CNNs) in processing data from different modalities, we propose a multimodal deep learning model for image popularity prediction. The proposed model exploits two dedicated CNNs to separately learn discriminative representations from the adopted input features and then efficiently merged them into a united network to predict the popularity.

DOI:10.6814/NCCU202100261

Chapter 3

Comparative Deep Learning Framework for Facial Age Estimation

3.1 Overview

In recent years, the development of automatic facial age estimation algorithms that demonstrate comparable or superior performances than that of the human ability of age estimation has become an attractive yet challenging topic. Conventional methods estimate the age of a person directly from the given facial image. In contrast, motivated by human cognitive processes, we propose a comparative deep learning framework, called comparative region convolutional neural network (CRCNN); this framework first compares the input face with reference faces of known age to generate a set of hints (comparative relations, i.e., whether the input face is younger or older than each reference), and then, all hints are aggregated to

DOI:10.6814/NCCU202100261

estimate the age of a person. Our approach has several advantages: (i) the age estimation task is split into several comparative stages, which is simpler compared to directly computing the person's age; (ii) in addition to the input face, side information (comparative relations) can be explicitly employed to benefit the estimation task; and (iii) a few incorrect comparisons do not considerably influence the accuracy of result, which makes this approach more robust than the conventional one. To the best of our knowledge, the proposed approach is the first comparative deep learning framework for facial age estimation. In addition, we proposed incorporating the method of auxiliary coordinates (MAC) for training, which reduces the ill-conditioning problem of the deep network and affords an efficient and distributed optimization. The experimental results on the FG-NET, MORPH, and IoG databases demonstrate that the proposed CRCNN model achieves a significant outperformance compared to the state-of-the-art methods, with a relative improvement of 13.24% (on FG-NET), 23.20% (on MORPH) in term of mean absolute error, and 4.74% (on IoG) in term of age group classification accuracy. The content of this chapter have been published in [110].

3.2 Introduction

The appearance of a human face changes with age. Therefore, facial appearance is a very important trait when estimating the age of a person, and facial age estimation is an essential component in several mobile and social media applications [115–120]. However, the age estimation by humans is not as easy a task as determining other facial information such as identity, expression, and gender. Hence, developing automatic facial age estimation

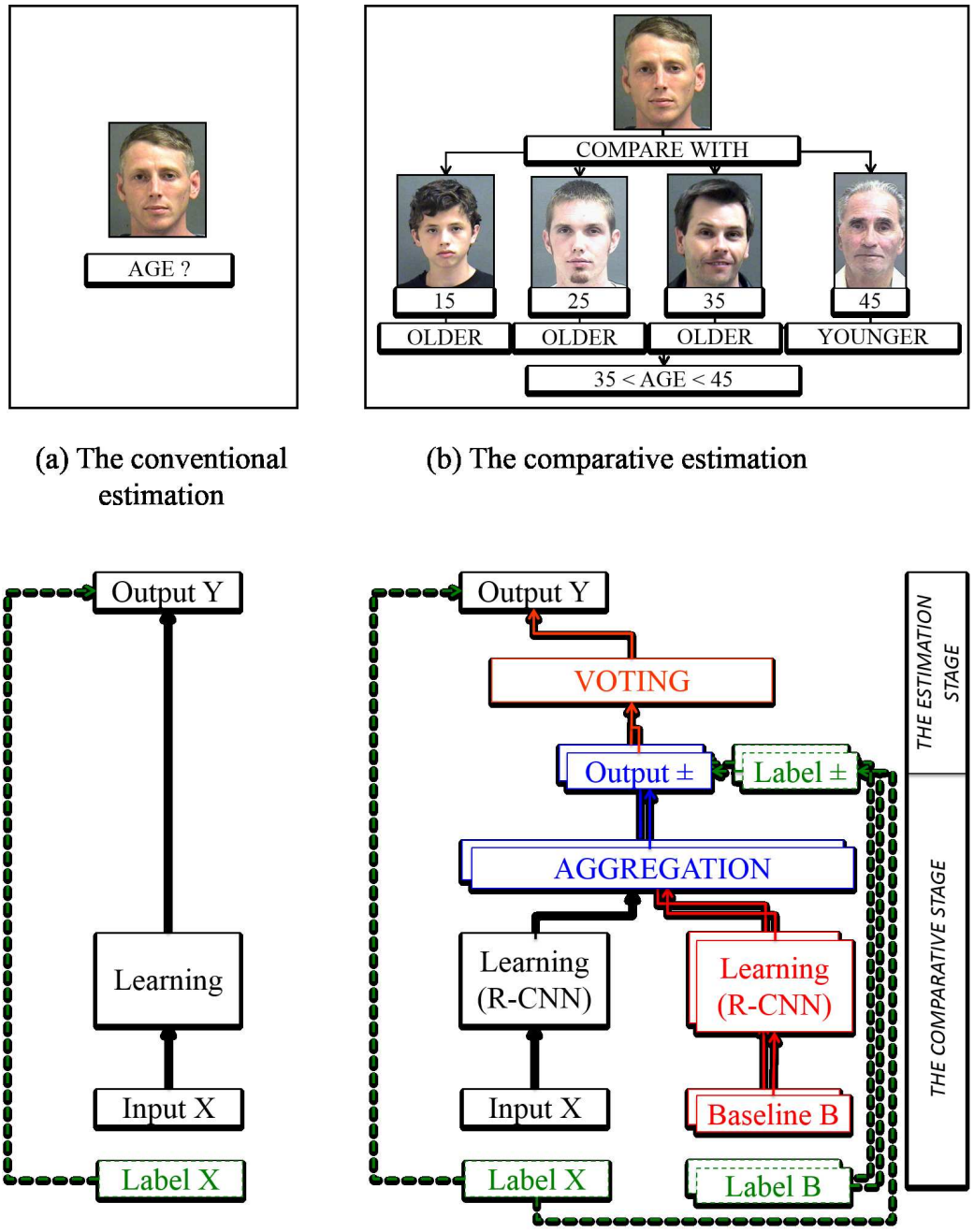
DOI:10.6814/NCCU202100261

methods that are comparable or superior to the human ability of age estimation has become an attractive yet challenging topic in recent years [46, 66, 64, 90, 121].

In the literature, the conventional way for facial age estimation is a direct method to estimate the age of a person by analysing his/her facial information (e.g. eyes, nose and so forth) directly from the facial image of the person, cf. Figs 3.1(a) and 3.1(c). In particular, only the input image is used to estimate the age of a person; however, estimating someone's precise age at a glance without any reference information is difficult, even for humans [90]. In response to the above challenges, our idea is to develop a facial age estimation algorithm inspired by the human cognitive processes [37]. In practice, humans commonly use several judgments to estimate a person's age, cf. Fig 3.1(b). First, they learn to establish connections between a known age and the corresponding facial cues of a person (direct method), and second, they employ the learned knowledge as a reference to evaluate whether an unseen face is younger or older than the reference (comparative method). The larger the number of available references, the more precise is the estimation of the age of an unseen face.

Therefore, a general mathematical framework, namely *comparative region-convolutional neural network (CRCNN)*, is proposed for facial age estimation, cf. Fig 3.1(d). Conceptually, we compare an unseen face with a set of selected references (labelled baseline samples) to determine if the person of the unseen face is younger or older than each of the baseline persons. We couple this comparative scheme with a specific deep learning architecture called region-convolutional neural network (R-CNN) [122]. The R-CNN is exploited to extract the most "iconic" local region from each facial image, where the spatial context (geometrical interrelation) of the extracted local regions can be also accounted for robust classification. In

DOI:10.6814/NCCU202100261



(a) The conventional estimation

(b) The comparative estimation

(c) The conventional paradigm

(d) The proposed comparative paradigm

Fig. 3.1 Schematic diagram of (a,c) the conventional paradigm for facial age estimation by learning the age information from a facial image directly, and (b,d) the proposed paradigm by aggregating the comparisons of a facial image with baseline samples to determine the age in a comparative manner.

DOI:10.6814/NCCU202100261

the proposed CRCNN framework, not only the input image is used, but also several other reference images are used as baseline samples to be compared with the input image. The comparison is equivalent to estimating whether the input person is younger or older than the others. Compared to the conventional paradigm, the proposed approach allows reformulating the estimation task into sequentially independent sub-problems, wherein each sub-problem represents a comparison (younger/older decision) between two images, which is considerably simpler than the initial task, i.e., estimating the exact age of an observed face. Further, by simply increasing the number of baseline samples, more side information (comparisons) can be exploited to benefit the estimation task, which can help achieve a more robust estimation. Finally, another advantage is that few incorrect comparisons do not significantly influence the accuracy of the age estimation because of leveraging many baseline samples.

Further, the traditional way to learn the parameters of a deep architecture is to minimize an objective function by computing the gradient over all the parameters using the backpropagation algorithm [123] with a nonlinear optimizer. However, the deep learning method is very difficult to train, especially because of the ill-conditioning problem and local minima issue [36]. These difficulties also complicate the manual tuning of deep learning parameters and convergence. In this study, we propose incorporating the recent method of auxiliary coordinates (MAC) [124] into our framework for training, which is an interesting direction toward the more efficient training of deep architecture. The method introduces a set of variables that can break the objective function dependency, thereby making the problem considerably better conditioned without nesting, and thus affording an efficient and distributed optimization. The contributions of this chapter can be summarized as follows:

DOI:10.6814/NCCU202100261

- To the best of our knowledge, our CRCNN framework is the first comparative deep learning approach for facial age estimation and has demonstrated that it can outperform state-of-the-art methods by experimenting with well-known face datasets.

- Instead of using the classical deep learning techniques such as a convolutional neural network (CNN) [111], we proposed the use of R-CNN to account for the spatial context of facial regions. Further, we improved the training efficiency of the deep architecture by incorporating the MAC technique; thus, the notorious ill-conditioning problem of deep learning can be alleviated.

- We implemented our mathematical framework with CAFFE [125], which is a popular deep learning platform that exploits the parallelization over multiple GPUs. The compatibility with CAFFE makes all the components of our mathematical implementation readily available to other researchers.

- The sensitivity of deep learning parameters makes it a non-trivial task to obtain an appropriate setting, and therefore, the systematic investigation on parametric optimization provides guidance to users who plan to extend our approach for future research.

3.3 Proposed Method: CRCNN Framework

The proposed *comparative region-convolutional neural network (CRCNN)* is a general mathematical framework for facial age estimation. It is developed by comparing an input face with a number of baseline samples to determine its age. The input face is compared

DOI:10.6814/NCCU202100261

with each baseline sample to determine whether the input face is older or younger than the baseline person. As a result, a set of hints (comparative relations) is generated. The estimation stage then aggregates the obtained set of hints to estimate the age of the input person. In Section 3.3.1, we first explain some of the preliminary definitions. Then, we provide an overview of our CRCNN framework in Section 3.3.2. Finally, each algorithmic component in our approach is explained in detail in Section 3.3.3.

3.3.1 Preliminary Definitions

Before explaining our CRCNN framework, we first define two terminologies: the baseline and the set of hints.

a) Baseline: The objective is to compare the age of an input image with those of a set of reference images, where the ages of these references are known. We define these references as the baseline. A baseline is composed of a set of reference samples, as many as possible to thoroughly cover the value range of the possible ages (e.g., labels). In other words, each baseline sample represents an age label. At the minimum, we use one baseline sample per label; and therefore, if we have M labels, we have M baseline samples in total. If we have K baseline samples per label, we have a total of MK baseline samples.

b) Set of hints: To understand the exploitation of the set of hints, we employ the example shown in Figure 3.2. To estimate the age of an input $\mathbf{X}$ (ground-truth age is 62), we first compare the input with the baseline samples $\mathbf{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_5\}$. A hint can be categorized

DOI:10.6814/NCCU202100261

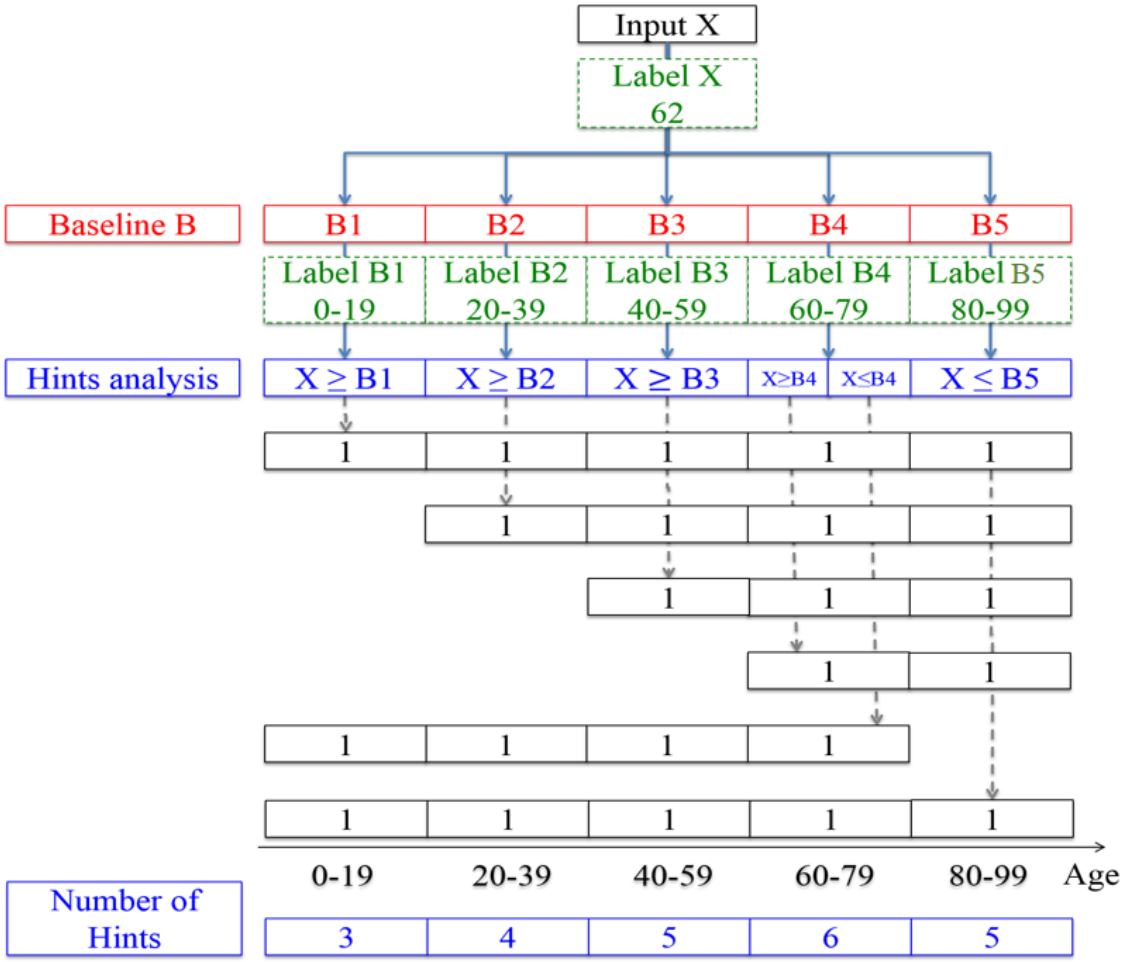


Fig. 3.2 Generation of a set of the hints (for simplicity, five labels are employed).

into "younger" or "older."[1] For each baseline sample, if the age of the input is estimated to be larger than the age of the baseline sample (i.e., the input person is estimated to be "older" than the baseline sample), we add a hint for the corresponding label of every baseline sample with an age larger than (or equal to) the comparing one. For example, we consider the comparison between $\mathbf{X}$ and $\mathbf{B}_2$. Since $\mathbf{X}$ is older than $\mathbf{B}_2$, we add a hint for the labels of

[1]Note that, in this study, the comparative relations of "younger" and "older" are actually defined to be "younger than or equal to" and "older than or equal to", respectively. The "same age" relation thus exists when the two relations hold simultaneously.

DOI:10.6814/NCCU202100261

$\mathbf{B}_2$, $\mathbf{B}_3$, $\mathbf{B}_4$, and $\mathbf{B}_5$ to indicate that they are all possible labels for $\mathbf{X}$. Similarly, if the input person is estimated to be "younger" than the baseline person, then we add a hint for the corresponding label of every baseline sample with its age smaller than (or equal to) the one being compared. Thus, the obtained hints of each label in the number are proportional to the likelihood that a label is the true label to the input. For example, as shown in Figure 3.2, $\mathbf{B}_4$ is the most likely label to $\mathbf{X}$ and $\mathbf{B}_1$ is the most unlikely one.

3.3.2 Overview of Our CRCNN Framework

Our CRCNN framework can be decomposed into two main stages, as shown in Figure 3.1(d).

Comparative stage (Collecting the hints)

After building a baseline, the input image is compared with each baseline sample. We use the R-CNN deep architecture to extract facial information from the images and then apply an energy function-based aggregation to generate the comparisons. Therefore, a set of hints was collected, wherein each hint represents a comparative relation (younger or older) that provides information to compute the estimated age at the next stage.

Estimation stage (Voting the hints)

This stage votes on the results from the set of hints to compute the estimated age.

DOI:10.6814/NCCU202100261

3.3.3 CRCNN Formulations

Considering $\mathscr{I}$ as a universal set of facial images and $\mathscr{L}$ be the corresponding label set of possible ages of a human being, we are given a training set of N facial images $\mathbf{X} \in \mathscr{I}$ and its label $\mathbf{Y} \in \mathscr{L}$. Let F denotes the deep architecture function. Instead of computing $\mathbf{Y}$ with F, as usual in the conventional paradigm:

$$\begin{array}{ccc} F : \mathscr{I} & \rightarrow & \mathscr{L} \\ \mathbf{X} & \mapsto & \mathbf{Y} = F(\mathbf{X}). \end{array} \tag{3.1}$$

The idea is to introduce a baseline $\mathbf{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_M\}$ from $\mathscr{I}$ with a composition function Ψ and Φ to decompose the task into two main parts. Note that $\mathbf{X}$ and $\mathbf{B}$ are usually disjoint. First, in the comparative stage, the comparison of $\mathbf{X}$ and the baseline $\mathbf{B}$ with Ψ provides the set of hints $\mathscr{H}$. Second, in the estimation stage, the vote of hints from the set of hints $\mathscr{H}$ is to obtain the final label $\mathscr{L}$ with Φ. Therefore, the proposed CRCNN approach is formulated as:

$$\begin{array}{ccccccc} (\mathscr{I} \times \mathscr{I}) & \underrightarrow{\Psi} & \mathscr{H} & \underrightarrow{\Phi} & \mathscr{L} \\ (\mathbf{X}, \mathbf{B}) & \mapsto & \mathbf{Z} = \Psi(\mathbf{X}, \mathbf{B}) & \mapsto & \mathbf{Y} = \Phi(\mathbf{Z}). \end{array}$$

Comparative stage

The set of hints $\mathbf{Z} \in \mathscr{H}$ is computed from $\mathbf{X} \in \mathscr{I}$ and $\mathbf{B} \in \mathscr{I}$ with the function Ψ, which is decomposed into

DOI:10.6814/NCCU202100261

$$\Psi = \Psi^R \circ \Psi^C \circ \Psi^L \circ \Psi^F \circ \Psi^A.$$

The first operator Ψ^R detects all regions where the facial information is selected by R-CNN to be the most relevant. The second operator Ψ^C is the convolutional step (including sub-sampling layers) that extracts a fixed-length feature vector from each region. The third and fourth operators (Ψ^L and Ψ^F) are locally and fully connected steps [111]. Finally, the features of both the input image and baseline samples are aggregated into the last operator Ψ^A, where an energy function approximates the age comparison with a distance metric.

Region-detection layer: Consider $\mathbf{X}_i \in \mathscr{I}$, an input image, a set of candidate regions $\{X_{i,j}\}_{j=1...J}$ is detected from $\mathbf{X}_i$ in order to extract more efficient facial information features. Each region $X_{i,j}$ is detected by the algorithm in [122]. The same region-detection operator Ψ^R is applied to each baseline sample $\mathbf{B}_m$, providing a set of candidate regions $\{B_{m,j'}\}_{j'=1...J'}$. Therefore, we denote the first hidden layer of our deep architecture by $\mathbf{H}_1$, which is formed with the region-detection layer. If no region detection is used (Ψ^R is equivalent to an identical function), then we set the output as the input image itself ($\{X_i\} = \{\mathbf{X}_i\}$).

Convolutional layers: The convolutional operator Ψ^C extracts features from the first hidden layer $\mathbf{H}_1$. Specifically, features are computed by forward propagation through a convolutional structure of $|C|$ layers with

$$\Psi^C = \Psi^C_1 \circ \Psi^C_2 \circ \cdots \circ \Psi^C_{|C|}.$$

DOI:10.6814/NCCU202100261

These steps expand the input into a set of simple local features. We denote $\mathbf{H}_k = \Psi_k^C(\mathbf{H}_{k-1})$ as the output of a convolutional layer for $k = 2,3,\ldots,|C|+1$. Further details of the convolutional layer are provided in [111]. We interpret these convolutional steps as an adaptive pre-processing step. The purpose of these convolutional steps is to extract low-level features such as simple edges and textures. Sub-sampling layers make the output of convolution networks more robust to local translations and small registrational errors, which is important in facial recognition problems.

Locally connected layers: After extracting features with Ψ^C, applied independently to $\mathbf{X}_i$ and $\mathbf{B}_m$, we first combine locally extracted features through $|L|$ locally connected layers with

$$\Psi^L = \Psi_1^L \circ \Psi_2^L \circ \cdots \circ \Psi_{|L|}^L,$$

resulting in $\mathbf{H}_k = \Psi_k^L(\mathbf{H}_{k-1})$ for $k = |C|+2, |C|+3, \ldots, |C|+|L|+1$. Similar to convolutional deep learning, locally connected layers apply a filter bank, but every location in the feature map learns a different set of filters. For example, information from an area between the eyes and eyebrows is combined with the one between the nose and the mouth; however, the two pieces of information are processed differently in the convolutional operation.

Fully connected layers: The fully connected operation Ψ^F computes all weights together with

DOI:10.6814/NCCU202100261

$$\Psi^F = \Psi_1^F \circ \Psi_2^F \circ \cdots \circ \Psi_{|F|}^F$$

and $\mathbf{H}_k = \Psi_k^F(\mathbf{H}_{k-1})$ for $k = |C|+|L|+2, |C|+|L|+3, \ldots, |C|+|L|+|F|+1$. Unlike in the locally connected operation where the inputs are locally combined, each output unit in the fully connected layers is connected to all inputs. Further, these layers can capture correlations between features captured in distant parts of the facial images, for example, the position and shape of the eyes and the position and shape of the mouth.

Aggregation: An EBM energy function [126] is exploited to aggregate both information regarding $\mathbf{X}_i$ and $\mathbf{B}_m$ from the fully connected operation to estimate if $\mathbf{X}_i$ is younger or older than $\mathbf{B}_m$. The advantage of the adopted energy function is that there is no need to estimate the normalized probability distributions over the input space. The scalar energy function E measures the compatibility between $\mathbf{X}_i$ and $\mathbf{B}_m$, and it leads to a set of hints associated with the in-between comparative relationship, as shown in Figure 3.2. This real-valued energy function is thus defined as $E(\mathbf{X}_i, \mathbf{B}_m) = ||G_W(\mathbf{X}_i) - G_W(\mathbf{B}_m)||$, where G_W denotes a mapping (subject to learning) to produce output vectors that are nearby for images from the same person and far away for images from different persons [126].

Learning is then performed by finding the deep architecture parameters that minimize a suitably designed loss function evaluated over a training set. Let L^- (or L^+) be the partial loss function if $\mathbf{X}_i$ is younger (or older) than $\mathbf{B}_m$; then, our loss function is

DOI:10.6814/NCCU202100261

$$L = (1 - \bar{\mathbf{Z}}_l)L^-(E(\mathbf{X}_i, \mathbf{B}_m)) + (\bar{\mathbf{Z}}_l)L^+(E(\mathbf{X}_i, \mathbf{B}_m)),$$

where $\bar{\mathbf{Z}}_l$ denotes the ground truth of the hint $\mathbf{Z}_l$. The partial loss function L^- (or L^+) is designed such that the minimization of L will decrease (or increase) the energy when $\mathbf{X}$ is younger (or older) than $\mathbf{B}_i$. A simple approach to achieve this is to make L^- monotonically decreasing, and L^+ monotonically increasing.

Estimation stage

Once the set of hints has been generated, the estimation stage is applied to vote based on the output information of the previous comparative stage to determine the age of the person. The representation of the set of hints in Figure 3.2 includes the number of hints for each label. This result is computed by applying a summation at each label. Therefore, the age of the input person can be estimated by considering the label with the most votes in a naive manner. In practice, to avoid the case where most votes appear in more than one label, we choose to use the real value outputted from the energy function E instead of the number of hints $\mathbf{Z}_i$ because the confidence of a vote is also embedded. That is, a larger value indicates higher confidence of a vote and vice versa.

3.3.4 Learning Method for the Comparative Stage

In this study, we propose incorporating the recent method of auxiliary coordinates (MAC) [124] for training the comparative stage. The MAC method decouples the typical learning

DOI:10.6814/NCCU202100261

problem of the comparative stage, which typically has an objective function of the form

$$\min \| \mathbf{Z} - \Psi(\mathbf{X}, \mathbf{B}) \|^2$$

into

$$\min \quad \| \mathbf{H}_{k+1} - \Psi_{k+1}(\mathbf{H}_k) \|^2$$

$$\min \quad \| \mathbf{Z} - \Psi_K(\mathbf{H}_K) \|^2$$

for $k = 1, 2, \ldots, |C| + |L| + |F|$ and $K = |C| + |L| + |F| + 1$. Note that MAC is applied only to the convolutional, locally, and fully connected layers such that $\Psi_k \in \{\Psi_k^C, \Psi_k^L, \Psi_k^F\}$. The problem becomes a set of small, independent minimization subproblems, each of which can be easily solved without back-propagating any gradients. The objective function is optimized over the hidden layer $\mathbf{H}$ and over the weights $\mathbf{W}$ (of the function Ψ) with the two functions below alternatively.

$$\mathbf{H}_{k-1} \xrightarrow{\Psi_{k-1}} \mathbf{H}_k \xrightarrow{\Psi_k} \mathbf{H}_{k+1}$$

$$\mathbf{H}_k \xrightarrow{\Psi_k} \mathbf{H}_{k+1}$$

DOI:10.6814/NCCU202100261

Specifically, optimizing the objective function over the hidden layer $\mathbf{H}_k$ implies optimizing the following nonlinear least-squares regression:

$$\min_{\mathbf{H}_k} \| \mathbf{H}_k - \Psi_{k-1}(\mathbf{H}_{k-1}) \|^2 + \| \mathbf{H}_{k+1} - \Psi_k(\mathbf{H}_k) \|^2$$

and alternatively, optimizing it over the weight $\mathbf{W}^k$ (of the function Ψ_k) with

$$\min_{\mathbf{W}^k} \| \mathbf{H}_{k+1} - \Psi_k(\mathbf{W}^k, \mathbf{H}_k) \|^2 .$$

Optimizing over the hidden layer $\mathbf{H}_k$ has fixed weights $\mathbf{W}^k$, and optimizing over the weight $\mathbf{W}^k$ has a fixed hidden layer $\mathbf{H}_k$. This minimization problem results in several independent, single-layer, single-unit problems that can be solved with existing algorithms without incurring extra programming cost. We solve this nonlinear least-squares fitting problem with a Gauss-Newton approach [127].

3.4 Experimental Results and Discussions

In this section, we present the results from a series of experiments designed to optimize and test the effectiveness of our CRCNN framework. We implemented our experiments using CAFFE in a machine with an Intel CPU 3.40-GHz dual-core processor. First, we present the general setting of our experiments. Second, we optimize the setting (i.e., we try our best to

DOI:10.6814/NCCU202100261

search for the best setting empirically) of our CRCNN approach. Finally, we compare our CRCNN approach with state-of-the-art methods in facial age estimation.

3.4.1 Experimental Setup

Datasets

We used three public datasets in the experiments, which are common benchmarks adopted in the related literature [90, 68, 128, 129]. The first is the FG-NET aging database [29]. There are 1,002 facial images from 82 subjects in this database. Each subject has 6–18 facial images at different ages, and each image is labelled by its real age. The ages are distributed over a wide range from 0–69 years. The dataset images exhibit large facial variations, such as significant changes in pose, illumination, and expression. The second dataset is the MORPH database [130]. There are 55,132 facial images from more than 13,000 subjects in this database. The average number of images per subject is 4. The ages of the facial images range from 16–77 years with a median age of 33 years. The faces are from different races, among which the African faces account for $\sim 77\%$; the European faces account for $\sim 19\%$; and Hispanic, Asian, Indian, and other races, the remaining 4%. Finally, the last dataset is the images of groups (IoG) dataset [131]. This dataset comprises 5,080 images with a total of 28,231 labeled faces. The images were acquired through searches on the photo-sharing website Flickr, and each face was assigned to one of the seven age groups: 0–2, 3–7, 8–12, 13–19, 20–36, 37–65, and 66+. As the images were collected from searches, there is an extremely uneven distribution of images across age and pose.

DOI:10.6814/NCCU202100261

Implementation Platform

CAFFE [125] is a BSD-licensed C++ library with Python and MATLAB bindings for training and deploying general-purpose convolutional neural networks and other deep models efficiently on commodity architectures. It is now a very popular deep learning platform, and we chose to implement our CRCNN framework based on it to give high extendibility for future practitioners to integrate their own implementations with our CRCNN framework.

Early and Late Fusion Schemes

We perform our mathematical comparative method with two different schemes: the early fusion and the late fusion [132]. The framework described in this study adopts the late fusion scheme, i.e., we extract features from the input image and each baseline sample separately and then fully connect all information into a final layer of the deep architecture. Alternately, the early fusion scheme first combines the input image with the baseline samples and then extracts information from both types of images simultaneously. Both fusion schemes are optimized, tested, and compared to the state-of-the-art results.

3.4.2 Optimization of Our CRCNN Framework

In this section, we present the optimization of our deep architecture to provide insights into the sensitivity of the parameters associated with our CRCNN framework. First, the performance of the comparative stage with different settings of the deep architecture parameters (e.g., fusion strategy, baseline, region detection, etc.) is shown in Figure 3.3. Each sub-figure

DOI:10.6814/NCCU202100261

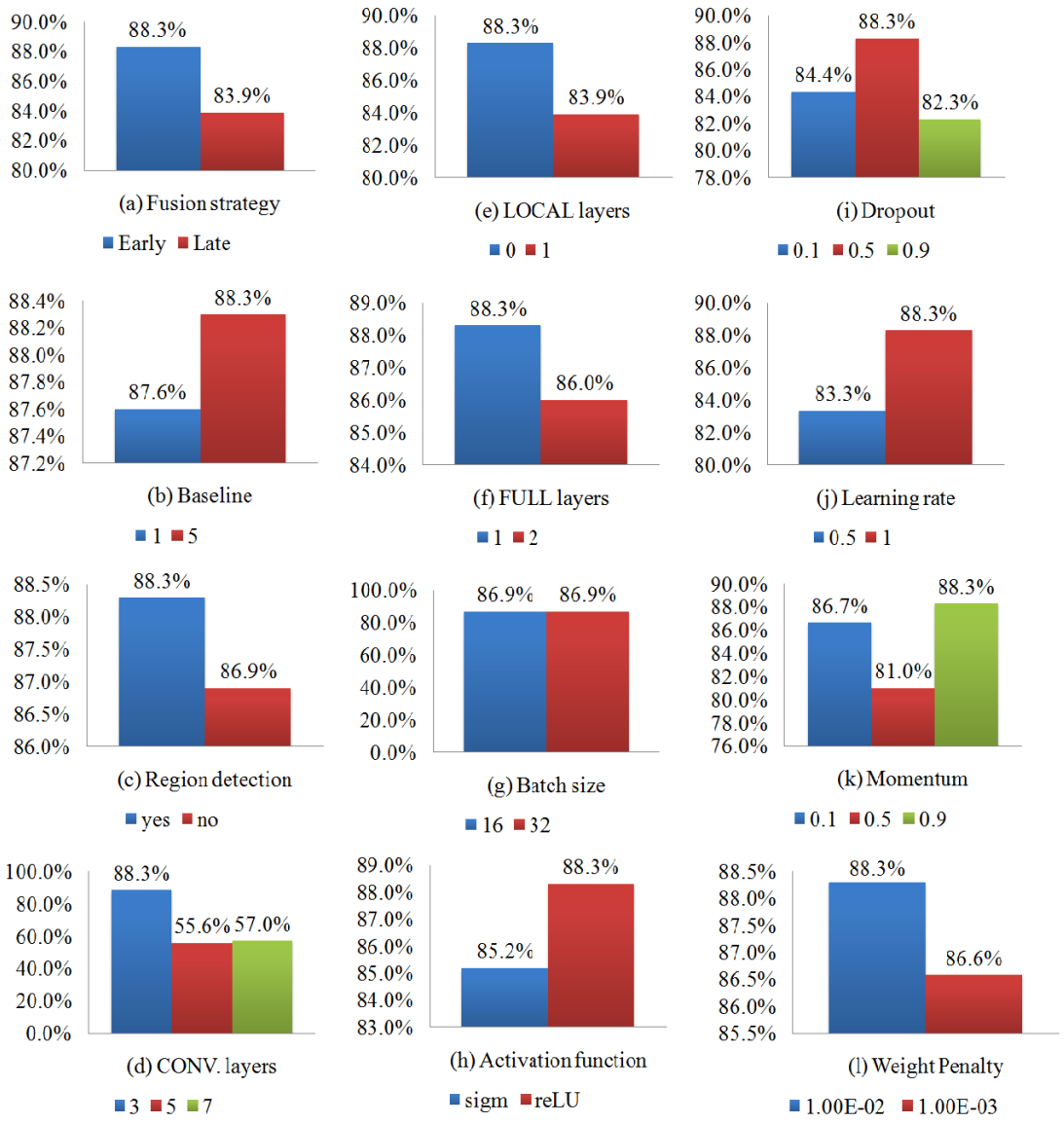


Fig. 3.3 Optimization of our CRCNN approach: Performance for the different settings of the deep architecture's parameters.

represents the performance of a parameter when in different values (or choices). Empirically optimal values of our CRCNN parameters obtained from the experiments are summarized in Table 3.1. Second, the sensitivity between the parameters is presented in Figure 3.4. Each sub-figure represents the correlation coefficient of a parameter and the others based on obtained

DOI:10.6814/NCCU202100261

Table 3.1 Optimized setting of our CRCNN method.

Deep architecture's parameters	**Optimized value**
Fusion	Early
Number of baseline samples	5
Region detection	Yes
Number of convolutional layers	3
Number of local-connected layers	0
Number of fully-connected layers	1
Batch size	32
Activation function	ReLU
Dropout	0.5
Learning rate	1
Momentum	0.9
Weight penalty	1e-2

performances (of the comparative stage). The lower the correlation coefficient is (close to 0), the more independent the two parameters are; the higher the correlation coefficient is (close to 1), the more dependency exists between them in terms of the performance of the comparative stage. For example, in Figure 3.4(g), the correlation coefficient of batch size (BS) and dropout (D) is less than 0.5 (weakly related), and the correlation coefficient of BS and itself is naturally 1 (perfectly related). Note that the raw image pixels are captured as the extracted features.

CRCNN Parameters

Fusion strategy (F): Early and late fusions are different in terms of weight sharing. In early fusion, both types of images (input and baseline) share the same set of weights, and in late fusion, each image has its own weight. As shown in Figure 3.3(a), the first value (88.3%) represents the accuracy when early fusion is applied to our CRCNN framework, and

DOI:10.6814/NCCU202100261

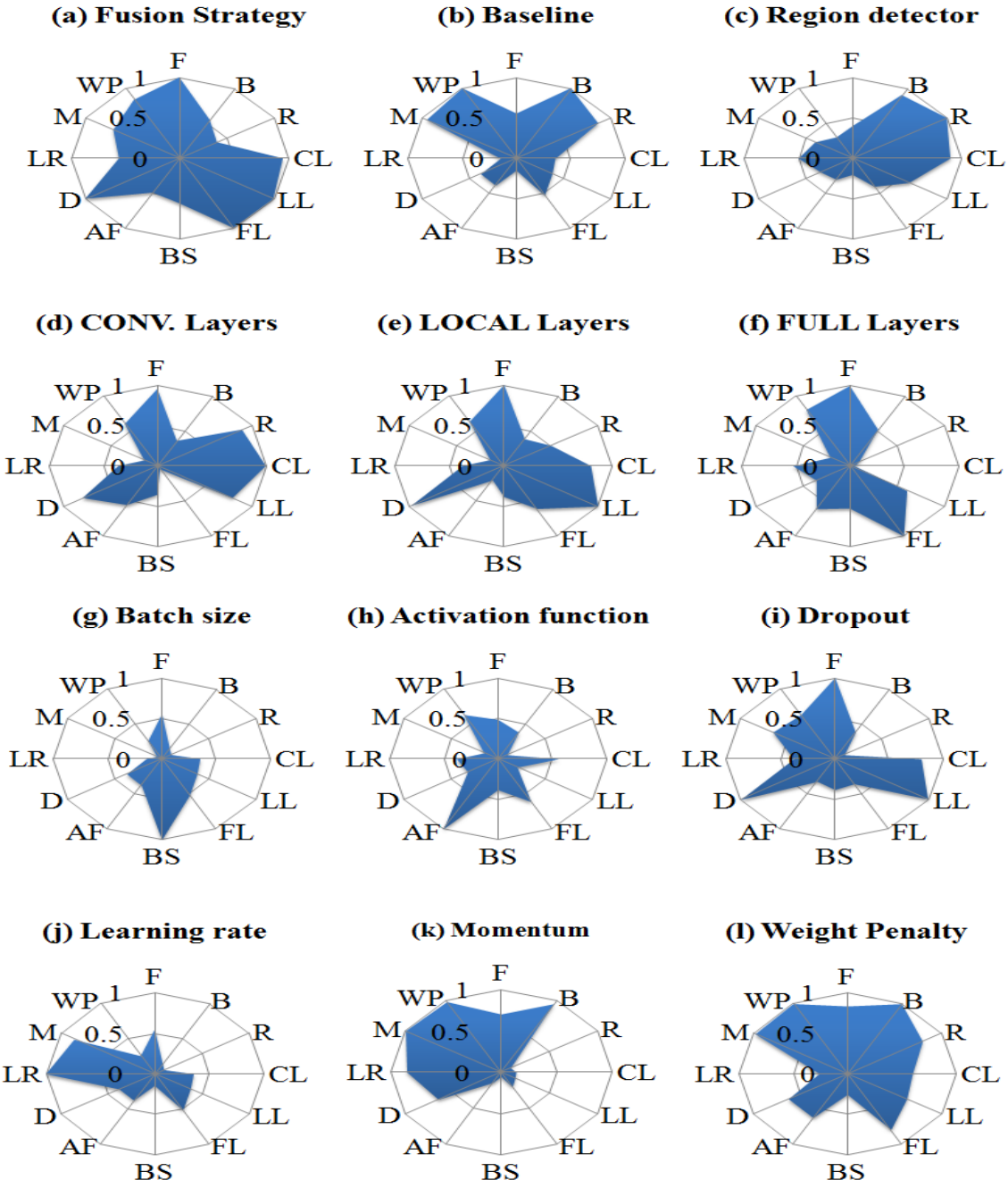


Fig. 3.4 Optimization of our CRCNN approach: Sensitivity of the deep architecture's parameters.

DOI:10.6814/NCCU202100261

the second value (83.9%) represents the accuracy when late fusion is applied. In other words, Figure 3.3(a) shows better accuracy when early fusion is applied. This observation intuitively corresponds to the fact that learning shared weights improves the inner relation between the input image and baseline. We observe in Figure 3.4(a) that the optimization of each fusion strategy depends on the entire deep architecture (i.e., convolutional layers, locally connected layers, and fully connected layers) and the value of the dropout.

Baseline (B): Each baseline sample is used as a reference to represent a range of possible ages (e.g., labels). In our optimization, we consider M baseline samples per label with $M = 1$ and 5. As expected and shown in Figure 3.3(b), a more robust computation is provided when $M > 1$ baseline samples represent each label. Correlations exist between this parameter and the region detection, and with several deep learning parameters, such as the momentum and the weight penalty (Figure 3.4(b)).

Region detection (R): We optimized our method with and without region detection. In other words, this optimization is equivalent to optimizing our CRCNN method by combining R-CNN [122] or classical CNN [111]. Figure 3.3(c) shows the results of this optimization, and it is clear that region detection Ψ^R can extract more robust features for improving performance. The performance of applying this detection depends on the setting of its input (e.g., baseline) and output (e.g., convolutional layers), as observed in Figure 3.4(c).

Convolutional layers (CL): We optimized the convolutional layers Ψ^C relating to the influence of the number of layers. Several numbers of layers have been experimented, and

DOI:10.6814/NCCU202100261

the results are shown in Figure 3.3(d). We observe that three convolutional layers provide the best results, and the number of layers is logically correlated with its previous and following layers (the region detector and the locally connected layer Ψ^C), and also with the value of dropout and the early/late fusion choice (Figure 3.4(d)).

Locally connected layers (LL): We optimized the locally connected layers Ψ^L. Figure 3.3(e) shows the results for a different numbers of layers. The most accurate result is provided when the convolutional layer Ψ^C is directly connected to the fully connected layer Ψ^F. Its influence on other parameters is the same as that of the convolutional layers (Figure 3.4(e)).

Fully connected layers (FL): The optimization of the fully connected layers Ψ^F is shown in Figure 3.3(f). We observe that only one fully connected layer is sufficient to yield the best results. Further, the optimization of the number of fully connected layers can be set independently (Figure 3.4(f)).

Batch size (BS): "Batch" learning accumulates contributions for all data points, and then updates the parameters. We use the "mini-batches" learning [133], where the parameters are updated after every n data points (i.e., this approach divides the dataset into piles and learns each pile separately). The computation time for learning the deep architecture depends on the number of epoches and the size of the batches. Figure 3.3(g) shows two different batch sizes. Empirically, we take `BS = 32`, and the batch size can be optimized independently (Figure 3.4(g)).

DOI:10.6814/NCCU202100261

Activation function (AF): The type of nonlinear activation function is typically chosen to be the logistic sigmoid `(sigm)` function or the rectified linear unit `reLU` function. We observe in Figure 3.3(h) that `reLU` has better accuracy than `sigm`. In general, `reLU` trains faster and outperforms the other activation functions. This parameter can also be set independently (Figure 3.4(h)).

Dropout (D): The dropout process is one wherein each hidden unit is randomly omitted from the deep architecture with a probability such that the hidden unit cannot rely on other hidden units being presented, based on the observation that this parameter is correlated with the deep architecture (Figure 3.4(i)). Previously, we observed the dependency between the influence of this parameter and the early/late fusion choice. Therefore, each fusion strategy leads to its own setting: `D = 0.5` for early fusion (Figure 3.3(i)) and `D = 0` for late fusion.

Learning rate (LR) and momentum (M): We continue the analysis with the learning rate and momentum. Each iteration updates the weight by the computed gradient. The learning rate indicates the convergence speed, and the momentum parameter introduces a *damping* effect on the search procedure, thereby avoiding oscillations in irregular areas of the error surface by averaging gradient components with opposite signs and accelerating the convergence in long flat areas. In our experiments, we observed that in Figures 3.3(j) and 3.3(k), the unit step and `M` both near to 1 converges better. Thus, we consider the `LR = 1` and `M = 0.9`, which are set dependently (Figures 3.4(j) and 3.4(k)). That is, the use of `M` in the age estimation task can help avoid the search procedure from being stopped at a

DOI:10.6814/NCCU202100261

local minimum, and it helps improve the convergence of the back propagation algorithm in general.

Weight penalty (WP): The last parameter is a constraint on the updating weight, and we observe in Figures 3.3(l) and 3.4(l) that the penalty can be set as `penalty = 1e-2`; further, this parameter will influence the setting of several parameters, such as the momentum, baseline, and fully connected layers.

In summary, the architecture of our CNN consists of three convolutional layers (CL), each of which is followed by rectification, max-pooling, and normalization; in addition, one fully connected layer (FL) is used. The network architecture is detailed as follows:

1. CL: The kernel size is 5×5, 1 stride - ReLU - pool 3×3, 2 stride - local response normalization (LRN).
2. CL: The kernel size is 5×5, 1 stride - ReLU - pool 3×3, 2 stride - LRN.
3. CL: The kernel size is 5×5, 1 stride - ReLU - pool 3×3, 2 stride - LRN.
4. FL.
5. SoftMax Loss Layer.

Computational cost

Given an input image, our comparative approach compares it with all k baseline samples; however, it does not compare it with all N training samples. For example, in our experiments,

DOI:10.6814/NCCU202100261

each age label is represented by one baseline sample, and we have nine labels, which makes k = 9. In other words, we only need to compute the comparative relationship of the input image k times, where k can be a small number and much less than N. Therefore, the computational cost of our approach is reasonable.

3.4.3 Discussions and Comparisons with State-of-the-art Methods

We compared our approach with other recent facial age estimation techniques, such as rKCCA [68], IIS-LLD [90], CPNN [90], OHRank [129], AGES [70], and two *aging function regression-based* methods WAS [29] and AAS [45]. Further, several conventional general-purpose classification methods, such as *k*-nearest neighbors (*k*NN) [134], back propagation neural network (BP) [135], C4.5 decision tree [136], SVM [137], adaptive network-based fuzzy inference system (ANFIS) [138], and ranking-based approaches, such as ranking SVM [96], RankBoost [97], and RankNet [98] were also used for the comparison. We trained our model using the popular leave-one-person-out (LOPO) test strategy [70], as suggested in the related benchmarks [90, 68, 129, 70]. In particular, we split the used datasets (FG-NET and MORPH) by adopting the same training/testing protocol for all comparison methods. For example, LOPO is used on the FG-NET dataset as follows: in each fold, images of one person are used as the testing set and those of the others are used as the training set. After 82 folds (the FG-NET dataset has a total of 82 subjects), each subject was used as the testing set, and in turn, the average results were computed from all estimates. However, because there are more than 13,000 subjects in the MORPH dataset, the LOPO test becomes too

DOI:10.6814/NCCU202100261

time-consuming. Therefore, we adopted 10-fold cross-validation instead for the MORPH dataset.

Our CRCNN method is configured with the deep learning parameters optimized in Section 3.4.2, and the results are detailed in Table 3.1. Here, the human tests are included for reference, which were performed on 5% samples from the FG-NET database and 60 samples from the MORPH database [90]. The performance of the age estimation was evaluated by the mean absolute error (MAE) metric. In statistics, the MAE is used to measure how close a prediction is to the ground truth. In our case, the MAE is the mean of absolute errors between estimates and true ages; i.e., MAE $= \sum_{k=1}^{N} |\widehat{a}_k - a_k|/N$, where $\widehat{a}_k$ and a_k denote the estimate and true ages of the sample image k, and N denotes the total number of samples. The standard deviations of the MORPH dataset are also listed in Table 3.2. For example, a number in the format $a \pm b$ means that the MAE a has a standard deviation of b. Some comparison methods (e.g., rKCCA and rKCCA+SVM) do not show standard deviations because they do not report the standard deviation values in the experiments of their works. For the results of the FG-NET dataset, we follow the common practice of the previous work (e.g., [90]) and do not indicate standard deviations. For example, as reported in [90], "the number of images for each person in the FG-NET database varies dramatically. Consequently, the standard deviation of the LOPO test on the FG-NET database becomes unstable". In other words, for the FG-NET database, the values of standard deviation are not statistically meaningful, and thus, these values are not shown. The statistics are listed in Table 3.2. As listed in the table, the best results (boldfaced) are obtained from our CRCNN approach (with the early fusion scheme). The second-best results are also from our CRCNN approach (with the late

DOI:10.6814/NCCU202100261

fusion scheme). Thus, the overall performance of the CRCNN is very encouraging; Our results are significantly better than those of all state-of-the-art methods. In comparison to the deep learning-based method, that is CPNN [90], we also achieved better performance with a relative improvement of 13.24% (from 4.76 to 4.13 on FG-NET) and 23.20% (from 4.87 to 3.74 on MORPH). These facts validate the robustness of the newly proposed comparative approach.

Table 3.2 Comparison with state-of-the-art methods on FG-NET and MORPH databases.

Method	**Database (FG-NET)**	**Database (MORPH)**
CRCNN (Early Fusion) (RCNN)	**4.13**	**3.74 ± 0.29**
CRCNN (Early Fusion) (CNN)	4.72	4.33 ± 0.27
CRCNN (Late Fusion) (RCNN)	4.20	3.81 ± 0.32
CRCNN (Late Fusion) (CNN)	4.81	4.52 ± 0.23
Ranking SVM [96]	5.24	6.49 ± 0.17
RankBoost [97]	5.67	6.83 ± 0.25
RankNet [98]	5.46	6.71 ± 0.24
rKCCA [68]	-	3.98
rKCCA + SVM [68]	-	3.92
IIS-LLD [90] (Gaussian)	5.77	5.67 ± 0.15
IIS-LLD [90](Triangle)	5.90	6.09 ± 0.14
IIS-LLD [90] (Single)	6.27	6.35 ± 0.17
CPNN [90](Gaussian)	4.76	4.87 ± 0.31
CPNN [90](Triangle)	5.07	4.91 ± 0.29
CPNN [90](Single)	5.31	6.59 ± 0.31
OHRank [129]	6.27	6.28 ± 0.18
AGES [70]	6.77	6.61 ± 0.11
WAS [29]	8.06	9.21 ± 0.16
AAS [45]	14.83	10.10 ± 0.26
kNN [134]	8.24	9.64 ± 0.24
BP [135]	11.85	12.59 ± 1.38
C4.5 [136]	9.34	7.48 ± 0.12
SVM [137]	7.25	7.34 ± 0.17
ANFIS [138]	8.86	9.24 ± 0.17
Human Tests (HumanA)	8.13	8.24
Human Tests (HumanB)	6.23	7.23

DOI:10.6814/NCCU202100261

We further performed an evaluation on the IoG database. This database consists of 28,231 facial images collected from Flicker. Each face is labeled under one of the seven age groups: 0–2, 3–7, 8–12, 13–19, 20–36, 37–65, and 66+. In our evaluation, we considered only faces with an interocular distance of more than 40 pixels, which resulted in a subset of 1,495 face images. Further, we reorganized the age labels into the child, teen, and adult classes with ages 0–12, 13–19, and 20+, respectively. The setting yielded the following number of samples per age group: 546, 250, and 699. Finally, we performed the same normalizations as in the previous experiments on all IoG faces. We compare our results with the ranking based methods, including [96–98], and the local binary pattern kernel density estimation (LBP-KDE) [128]. The age group classification performance is summarized in Table 3.3. We can observe the better performance of our approach over the state-of-the-art methods with a relative improvement from 4.74% (in LBP-KDE) to 13.74% (in RankBoost). We believe that our method outperforms other approaches, thereby demonstrating its effectiveness for practical applications.

Table 3.3 Comparison with state-of-the-art methods on IoG database.

Method	**Database (IoG)**
CRCNN (Early Fusion) (RCNN)	**66.41%**
CRCNN (Early Fusion) (CNN)	63.16%
CRCNN (Late Fusion) (RCNN)	65.48%
CRCNN (Late Fusion) (CNN)	62.19 %
LBP-KDE [128]	61.67%
Ranking SVM [96]	56.17%
RankBoost [97]	52.67%
RankNet [98]	55.08%

DOI:10.6814/NCCU202100261

3.5 Summary

In this chapter, we proposed a novel comparative deep learning framework for facial age estimation called comparative region convolutional neural network (CRCNN). Motivated by human cognitive processes, we used a comparative approach to determine the age of an unseen person. To the best of our knowledge, this is the first comparative approach in deep learning for facial age estimation. The experimental results validate the superior performance of our CRCNN approach over state-of-the-art methods.

DOI:10.6814/NCCU202100261

Chapter 4

Multimodal Deep Learning Framework for Image Popularity Prediction on Social Media

4.1 Overview

Billions of photos are uploaded to the web daily through various types of social networks. Some of these images receive millions of views and become popular, whereas others remain completely unnoticed. This raises the problem of predicting image popularity on social media. The popularity of an image can be affected by several factors, such as visual content, aesthetic quality, user, post metadata, and time. Thus, considering all these factors is essential for accurately predicting image popularity. In addition, the efficiency of the predictive

DOI:10.6814/NCCU202100261

model also plays a crucial role. In this chapter, motivated by multimodal learning, which uses information from various modalities, and the current success of convolutional neural networks (CNNs) in various fields, we propose a deep learning model, called visual-social convolutional neural network (VSCNN), which predicts the popularity of a posted image by incorporating various types of visual and social features into a unified network model. VSCNN first learns to extract high-level representations from the input visual and social features by utilizing two individual CNNs. The outputs of these two networks are then fused into a joint network to estimate the popularity score in the output layer. We assess the performance of the proposed method by conducting extensive experiments on a dataset of approximately 432K images posted on Flickr. The simulation results demonstrate that the proposed VSCNN model significantly outperforms state-of-the-art models, with a relative improvement of greater than 2.33%, 7.59%, and 14.16% in terms of Spearman's Rho, mean absolute error, and mean squared error, respectively. The content of this chapter have been published in [139].

4.2 Introduction

Social media websites (e.g., Flickr, Twitter, and Facebook) allow users to create and share content (e.g., by liking, commenting, or viewing). Consequently, social media platforms have become an inseparable part of our daily lives, with significant social content generated on these platforms. The explosive growth of social media content (i.e., texts, images, audios, and videos) and the interactive behavior between web users result in that only a

DOI:10.6814/NCCU202100261

small portion of online social content attracts significant attention and becomes popular, whereas its vast majority either receives little attention or is entirely overlooked. Therefore, extensive efforts have been expended in the past few years to predict social media content popularity, understand its variation, and evaluate its growth [140–145, 38]. This popularity reflects user interests and provides opportunities to understand user interaction with online content, as well as information diffusion through social media websites. Hence, an accurate popularity prediction of online content may improve user experience and service effectiveness. Moreover, it can significantly influence several important applications, such as online advertising [58, 59], information retrieval [60], online product marketing [146], and content recommendation [147].

Popularity prediction on social media is usually defined as the problem of estimating the rating scores, view counts, or click-through of a post [103]. In this study, image popularity prediction on social media websites is analyzed to better understand the popularity factors for a particular image. Although this problem has recently received significant attention [40, 39, 41, 99], it remains a challenging task. For example, image popularity prediction can be significantly influenced by various factors (and features), such as visual content, aesthetic quality, user, post metadata, and time; therefore, considering all this multimodal information is crucial for an efficient prediction. Moreover, it is nontrivial to select an appropriate model that can make better use of the various features contributing to image popularity and accurately predict it. For example, simple machine learning schemes (e.g., support vector and decision tree regression) learn to predict by being fed with highly structured data, thus requiring time and skill to fine-tune the hyperparameters. However, to obtain accurate

DOI:10.6814/NCCU202100261

prediction results, it is critical to construct a prediction model capable of learning through a more abstractive data representation and optimizing the extracted features.

Accordingly, we address the image popularity prediction problem by analyzing a large-scale dataset collected from Flickr to investigate two essential components that may contribute to the popularity of an image; namely, visual content and social context. In particular, we examine the effect of the visual content of an image on its popularity by adopting different types of features that describe various visual aspects of the image, including high-level, low-level, and deep learning features. These are extracted by applying several techniques from machine learning and computer vision. Additionally, we explore the significant role of social context information associated with images and their owners by analyzing the following three types of social features: user, post metadata, and time. To demonstrate the efficacy of the proposed features, we propose a computational deep learning model, called visual-social convolutional neural network (VSCNN), which uses two individual CNNs to learn high-level representations of the visual and social features independently. The outputs of the two networks are then merged into a shared network to learn joint multimodal features and compute the popularity score in the output layer. End-to-end learning is employed to train the entire model, and the weights of its parameters are learned through back-propagation. In a nutshell, the contribution of this chapter can be summarized as follows:

- We demonstrate a comprehensive exploration of the independent benefits and predictive power of various types of visual and social context features towards the popularity of

DOI:10.6814/NCCU202100261

an image. We further demonstrate that these multimodal features can be combined effectively to enhance prediction performance.

- We propose a deep learning VSCNN model for predicting image popularity on social media. VSCNN uses dedicated CNNs to learn structural and discriminative representations from the input visual and social features, achieving considerable performance in predicting image popularity compared with several traditional machine learning schemes.
- We effectively set the architectures and parameters of the adopted CNNs to fit the multimodal information, i.e., social information and visual content of the image.
- We demonstrate that processing visual and social features using the late fusion scheme is significantly better than using the early fusion scheme.
- We use a large-scale dataset of approximately 432K images posted on Flickr to evaluate the performance of the proposed VSCNN model. The simulation results demonstrate that VSCNN achieves competitive performance and outperforms six baseline models and other state-of-the-art methods.

4.3 Features

In the literature, many types of features have been proposed for image popularity prediction [38, 99]. In this section, we analyze various kinds of features that can influence image popularity. First, in Section 4.3.1, we investigate some visual features that could be used

DOI:10.6814/NCCU202100261

to describe different visual facets of images based on their content. Then, in Section 4.3.2, we explore several social features based on the contextual information of images and their owners.

4.3.1 Visual Content Features

No one can deny that an image's popularity is related to some extent to its visual content. The visual information represents the most interesting parts of the image. These parts are commonly containing the salient objects in the image. To demonstrate the influence of an image's content on its popularity, we adopt different types of visual features, including low-level, high-level, and deep learning features. More details about these features are presented as described below.

Low-level Features

The low-level features describe the visual content of the image, which can be automatically extracted from the pixel information. There are several types of low-level computer vision features, such as texture, color, shape, gist, gradient, and spatial location. These features are likely used by humans for visual processing. In this study, we adopt the three following features: color, texture, and gist. For each of the features, we describe our motivation and the method used for extraction below.

Color: A perfect color distribution in an image attracts viewer attention and aids in determining object properties and understanding scenes. In this study, a color histogram

DOI:10.6814/NCCU202100261

descriptor that results in a vector of 32 dimensions and characterizes the color feature is used [148].

Texture: We routinely interact both visually and through touch with various textures and materials in our surroundings. Therefore, individuals commonly have a different visual perception of images with different texture features. The texture feature is often used to describe the homogeneity of colors or intensities in an image. It can also be used to identify the most interesting objects or regions [149]. To investigate the importance of this type of feature in predicting image popularity, we employ one of the most widely used features for texture description, namely, local binary patterns (LBP) [150]. More precisely, we use the uniform LBP descriptor [151], resulting in a 59-dimensional feature vector.

Gist: The Gist descriptor has demonstrated high performance in several tasks of computer vision such as scene classification. It provides a rough description of a scene by epitomizing the gradient information (scales and orientations) for various parts of a photo. To extract the GIST feature of an image, we adopt the widely used GIST descriptor proposed in [152], resulting in a feature vector with 512 dimensions.

High-level Features

The quality and aesthetic appearance of an image are important for its popularity. For instance, the clear images, as well as images that contain appealing or interesting objects, usually attract significant viewer attention and become popular. Therefore, we adopt certain aesthetic features for image popularity prediction based on the various photographic techniques and

DOI:10.6814/NCCU202100261

the aesthetic standards used by professional photographers. These features are designed to evaluate the visual quality of a photograph by separating the subject area from the background using the blur detection technique [153]. Then, based on the result of this separation process, six types of aesthetic features are computed as described below.

- **Clarity contrast:** Clarity contrast indicates a specific partition of an image that can be easily recognized because of its obvious difference from the background of the image. To attract the viewer's attention to the key point of a photograph and to isolate the subject region from the background, professional photographers normally adjust the lens to keep the subject in focus and make the background out of focus. Accordingly, a clear photograph will have relatively more high-frequency components than a blurred photograph [153]. To characterize this property, the clarity contrast feature is defined as:

$$f_c = \frac{\frac{\|M_S\|^2}{\|S\|^2}}{\frac{\|M_I\|^2}{\|I\|^2}}, \tag{4.1}$$

where $\| S \|^2$ and $\| I \|^2$ are the areas of the subject region and the original photograph, respectively, and

$$M_I = \big\{(m,n) \mid F_I(m,n) > \gamma \max\big\{F_I(m,n)\big\}\big\}, \tag{4.2}$$

$$M_S = \big\{(m,n) \mid F_S(m,n) > \gamma \max\big\{F_S(m,n)\big\}\big\}, \tag{4.3}$$

$$F_I = FFT(I),\ F_S = FFT(S). \tag{4.4}$$

DOI:10.6814/NCCU202100261

$\frac{\|M_S\|^2}{\|S\|^2}$ and $\frac{\|M_I\|^2}{\|I\|^2}$ are the ratios of the area of the high-frequency components to the area of all frequency components in S and I, respectively. m and n are the spatial frequencies. γ is a predefined threshold set to 0.2 in our experiments. $FFT(I)$ and $FFT(S)$ are the fast Fourier transforms calculated over I and S, respectively. The value of f_c is going to be high if the subject area of the given image I is in focus and the background is out of focus.

- **Hue count:** The hue count of an image is a metric of its simplicity. It can also be used to evaluate image quality. Although professional photographs appear bright and vivid, their hue number is normally less than that of amateur photographs. We thus compute the hue count feature of an image using a 20-bin color histogram H_c, which is computed on the good hue values. This can be formulated as follows [154]:

$$f_l = 20 - N_c, \tag{4.5}$$

$$N_c = \{i \mid H_c(i) > \beta m\}. \tag{4.6}$$

where N_c denotes the set of bins with values larger than βm, m is the maximum histogram value, and β is used to control the noise sensitivity of the hue count. We selected $\beta = 0.05$ in our experiments.

- **Brightness contrast:** Brightness contrast implies the difference in brightness of two adjacent surfaces. In high-quality photographs, the subject area's brightness significantly differs from that of the background because professional photographers

DOI:10.6814/NCCU202100261

frequently use different subject and background lightings. However, most amateurs use natural lighting and allow the camera to adjust the brightness of a picture automatically; this usually reduces the difference in brightness between the subject area and the background. To discern the difference between these two types of photographs, the brightness contrast feature f_b is calculated as [155]:

$$f_b = \ln\left(\frac{B_s}{B_b}\right), \tag{4.7}$$

where B_s and B_b denote the average brightness of the subject region and the background, respectively.

- **Color entropy:** Due to the unique interrelationship between the color planes of drawings and natural photos, the entropy of RGB and Lab color space components is calculated to distinguish drawings from natural photos [156].

- **Composition geometry:** The good geometrical composition is a fundamental demand to obtain high-quality photographs. The rule of thirds is one of the most important photographic composition principles utilized by professional photographers to bring more balance and high quality to their photos. If a photo is divided into nine parts of equal size by two equally-spaced horizontal lines and two equally-spaced vertical lines, the rule of thirds suggests that the subjects of the photo and any important compositional objects should be placed along these lines or their intersections. This is because most of the studies have demonstrated that when viewing images, people

DOI:10.6814/NCCU202100261

usually look at one of the intersection points rather than the center of the image. To formulate this criterion, the composition feature is defined as [155]:

$$f_m = \min_{j=1,2,3,4} \sqrt{\frac{\left(C_{s_x} - P_{j_x}\right)^2}{W^2} + \frac{\left(C_{s_y} - P_{j_y}\right)^2}{H^2}}. \tag{4.8}$$

where $\left(C_{s_x}, C_{s_y}\right)$ is the centroid of the subject area, $\left(P_{j_x}, P_{j_y}\right)$, $j = 1,2,3,4$, are the four intersection points in the photo, and W and H are the width and height of the photo.

- **Background simplicity:** Professional photographers normally maintain simplicity within the shot to enhance the composition of the photo. This is because photographs that are clean and free of distracting backgrounds look more appealing and naturally draw the attention of a viewer to the subject. The color distribution in a simple background tends to be less dispersed. Therefore, we compute the simplicity of the background using the color distribution of the background [157]. First, we consider the regions of an image not determined as a subject region to be the background. Then, each of the RGB channels of the image is quantized into 16 values, generating a histogram H_b of $16 \times 16 \times 16 = 4096$ bins, that shows the numbers of quantized colors present in the background. Thus, the feature that represents the background simplicity of an image can be calculated from H_b as [153]:

$$f_s = \left(\frac{N_b}{4096}\right) \times 100\%, \tag{4.9}$$

$$N_b = \left\{ i \mid H_b(i) \geq \alpha h_{max} \right\}. \tag{4.10}$$

DOI:10.6814/NCCU202100261

where h_{max} is the maximum count among all the bins of the histogram and α is used to control the noise sensitivity of the hue count. We choose $\alpha = 0.01$ in our experiments.

In this study, we combine the six aesthetic features indicated above, resulting in an 11-dimensional feature vector.

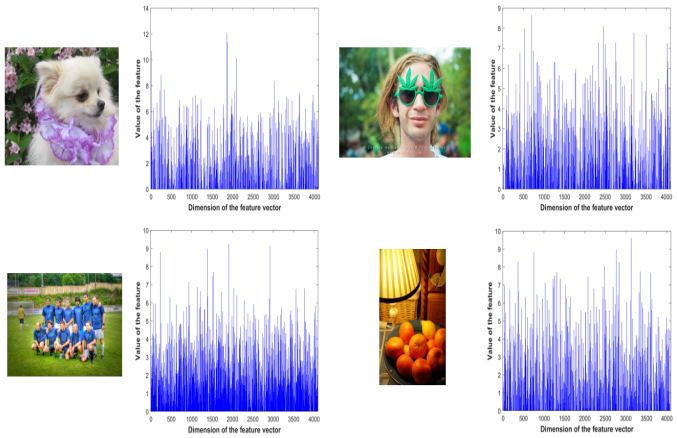


Fig. 4.1. The plots of deep learning feature vector values of different images from the dataset.

Deep Learning Features

Recently, deep learning methods have been widely used for image representation owing to their effectiveness [108], [111]. In this study, the CNN architecture of the VGG19 model was employed to learn the deep features of photographs [108]. The VGG19 model was trained on 1.2 million images from the ImageNet database to classify these images into 1000 categories [111]. The Keras framework of the VGG19 pre-trained CNN model [158] was used for feature extraction from the layer situated immediately prior to the final classification layer, (i.e., the last fully connected layer (fc7)). The output of this layer is a 4,096-dimensional

DOI:10.6814/NCCU202100261

feature vector. A few images selected from the dataset and the plots of their respective deep feature vector values are shown in Fig. 4.1.

4.3.2 Social Context Features

Previous studies demonstrated that the popularity of an image depends not only on its content, but also the social information regarding the user uploading the image, as well as the textual information associated with it [38, 99]. In this section, we attempt to determine the extent to which social features can influence image popularity. The following three types of social features are analyzed: user, post metadata, and time.

User

The popularity of a user highly correlates with the popularity of his/her posted images. Therefore, we adopted several user-centered features, which are listed below, and will have the same value for all photographs posted by the same user.

User id: This is defined as a unique integer number (ranging from 1 to 135) according to the average view count of all images of a user (i.e., a greater average number of views implies a larger value). Thus, this number is a unique identification of each user and can be used directly in the prediction model.

Average views: The average view count of all user-uploaded photos.

Group count: The number of groups to which a user subscribes.

DOI:10.6814/NCCU202100261

Member count: The mean number of members in the groups to which a user subscribes.

Image count: The total number of photographs posted by a user.

To characterize the effects of these features on predicting image popularity, we computed their rank correlation with the image popularity score using Spearman's rank correlation coefficient (Spearman's Rho) [159]. The value of the correlation coefficient ranges from [-1, 1], where a score of 1 (resp. -1) indicates an ideal positive (resp. negative) association, and a score of zero indicates no correlation. The results are shown in Table 4.1; both, user id and image views have a strong positive correlation with image popularity (Spearman's Rho = 0.74).

Table 4.1 Spearman's Rho values for the correlation of user features with popularity score.

Feature	Spearman's Rho
User id	**0.74**
Average views	**0.74**
Group count	0.067
Member count	0.19
Image count	- 0.35

Post Metadata

The contextual information associated with an uploaded image (e.g., tags, comments, or title) can also influence its popularity. For instance, an image with a large number of tags is expected to appear more frequently in search results. Therefore, we consider certain image contextual features for popularity prediction, which refer to image-related metadata, and most are entered by the user. The image-context features adopted in this study are listed below.

DOI:10.6814/NCCU202100261

Tag count: The number of tags annotated by a user on a posted photograph.

Title length: The number of characters in the title of a photograph.

Description length: The character count in the image description.

Tagged people: A binary number (1 or 0) indicating whether a given photograph has tagged people or not.

Comment count: The number of comments an image has obtained from other users.

We calculated the relationship between each of these features and the popularity of an image, as conducted for user features. The results of Spearman's Rho are listed in Table 4.2. Note, most of the post features have a significant positive correlation with popularity, except the tagged people feature, which has a slight positive correlation of 0.0043.

Table 4.2 Spearman's Rho values for the correlation of post metadata features with popularity score.

Feature	Spearman's Rho
Tag count	0.49
Title length	0.22
Description length	0.51
Tagged people	0.0043

Considering the results shown in Tables 4.1 and 4.2, the user id and image views features have the highest Spearman's Rho scores, which implies that user-centric features are the most effective in predicting image popularity. This also agrees with what we expected because popular users usually have a significant number of followers, indicating their images are more likely to receive a larger number of views after uploading on social media and thus become popular.

DOI:10.6814/NCCU202100261

Time

Along with the user and post features for predicting image popularity, there is a strong dependence on time features. For instance, users tend to become more active on social websites during the weekend. Thus, images posted at these time slots would naturally be expected to receive more views and therefore more ratings. Hence, we consider the following time features: post day, post month, post time, and post duration. The definitions of these features are as follows:

Post day: The day of the week on which a photograph is posted. We encoded the day number using one-hot encoding of a 7-dimensional vector.

Post month: The month in which a photograph is posted. We encoded the month number using one-hot encoding of a 12-dimensional vector.

Post time: The period of day during which a photograph is posted. The day was divided into four segments (six hours in each segment) assuming that the photograph is posted either in the morning (06:00 to 11:59), afternoon (12:00 to 17:59), evening (18:00 to 23:59), or night (00:00 to 05:59). Then, the post time was encoded using the one-hot encoding of a 4-dimensional vector.

Post duration: The amount of time in days during which the photograph remained posted on Flickr.

DOI:10.6814/NCCU202100261

4.4 Methodology

In this section, we explain the details of the proposed framework for predicting social media image popularity; moreover, we present a brief description of the baseline models.

4.4.1 Overview of Proposed Framework

The overall diagram of the proposed framework is shown in Fig. 4.2. It consists of the following two phases: feature extraction and VSCNN regression model. In the feature extraction phase, for each post in the dataset, we extract visual features from the image and social context features from its corresponding post-context information, as shown in Fig. 4.2 (a). The extracted visual features are integrated to obtain a final feature vector of 4,710 dimensions that describe the different visual facets of the image. Then, principal component analysis (PCA) [160] is performed to decrease the dimensionality of this vector from 4,710 to 20 and to select only the prevalent features. This results in a visual-PCA descriptor of 20 dimensions, which is denoted by X. Finally, the values of the X features are normalized so that all of them belong to the same scale. Similarly, the same procedure is applied to the corresponding extracted social features to obtain a normalized social-PCA descriptor of 14 dimensions, which is denoted by Z. The obtained X and Z are used as inputs to the proposed VSCNN model to predict the popularity of the corresponding post.

As shown in Fig. 4.2 (b), the proposed VSCNN model consists of two individual CNNs that are used to derive the structural and discriminative representations from the visual (X) and social (Z) features; namely the visual network and social network. Each of these

DOI:10.6814/NCCU202100261

networks is a one-dimensional CNN consisting of three convolutional layers. The rectified linear unit (ReLU) is employed as the activation function for each convolutional layer to avoid the vanishing gradient problem in the training phase. A fusion network is used to combine the outputs of these networks into a unified network. It consists of one merged layer and two fully connected layers. The merged layer is used to concatenate the outputs of the last convolutional layer of the visual network and that of the social network and generate the inputs of the first fully connected layer. Finally, the outputs of the second fully connected layer are summed at the final node, generating the predicted popularity score. In the diagram, the convolutional and fully-connected layers are denoted by Conv1D_v1, Conv1D_v2, Conv1D_v3, Conv1D_s1, Conv1D_s2, Conv1D_s3, FC1, and FC2, where the subscripts "v" and "s" indicate visual and social features, respectively.

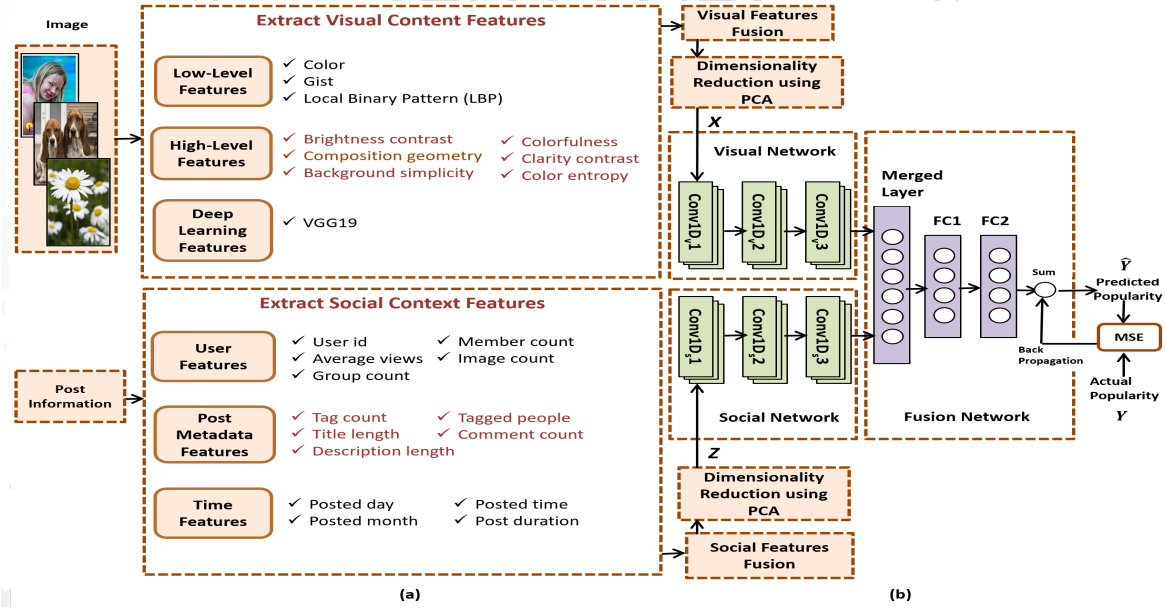


Fig. 4.2. Diagram of the proposed framework for image popularity prediction. (a) Feature extraction, and (b) Proposed VSCNN regression model.

DOI:10.6814/NCCU202100261

4.4.2 Training the VSCNN Model

Similar to the other CNNs, we first prepare a set of training samples (N) to train the VSCNN model. Each sample comprises of the posted image, post-context information, and corresponding popularity score (Y). We then calculate the visual feature descriptor (X) and the corresponding social feature descriptor (Z) for each sample, as indicated above. For each iteration, we obtain the output of the visual network as

$$V_i = Conv1D_v3\Big(Conv1D_v2\big(Conv1D_v1(X_i)\big)\Big),\ i = 1...N. \tag{4.11}$$

Similarly, the output of the social network is as follows:

$$S_i = Conv1D_s3\Big(Conv1D_s2\big(Conv1D_s1(Z_i)\big)\Big),\ i = 1...N. \tag{4.12}$$

Next, we flatten V_i and S_i, and concatenate the two feature vectors using a merge layer, and then we use the resulting concatenated feature vector as the input of the fusion network, $F_i = [V_i^{'}\ S_i^{'}]^{'}$. Thus, a fully connected cascade-feed-forward network can be calculated as follows:

$$\hat{Y}_i = FC2\Big(FC1\big(F_i\big)\Big),\ i = 1...N. \tag{4.13}$$

Let θ denote the parameters of the VSCNN model. First, they are initialized using random values ranging between -1 and 1, and then are trained by optimizing the following mean

DOI:10.6814/NCCU202100261

squared error (MSE) cost function using back-propagation:

$$MSE(\theta) = \min_{\theta} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \hat{Y}_i - Y_i \right\|_2^2 \right). \tag{4.14}$$

A stride of size 1 was adopted in the networks of the VSCNN model. To avoid overfitting, a dropout of 0.1 was adopted after each layer, except for the last fully connected layer, in which a dropout of 0.2 was used. Additionally, batch normalization was applied to each convolutional layer to increase the stability of the CNNs. Further details regarding the configuration of the VSCNN model are presented in Table 4.3.

Table 4.3 Configuration of the VSCNN Model.

Layer	Kernel	Activation Function	Numberof Neurons
$\text{Conv1D}_\text{V}1$	3	ReLU	32
$\text{Conv1D}_\text{V}2$	3	ReLU	64
$\text{Conv1D}_\text{V}3$	3	ReLU	128
$\text{Conv1D}_\text{S}1$	2	ReLU	32
$\text{Conv1D}_\text{S}2$	2	ReLU	64
$\text{Conv1D}_\text{S}3$	2	ReLU	128
Merged Layer			4736
FC1		ReLU	1024
FC2		ReLU	500

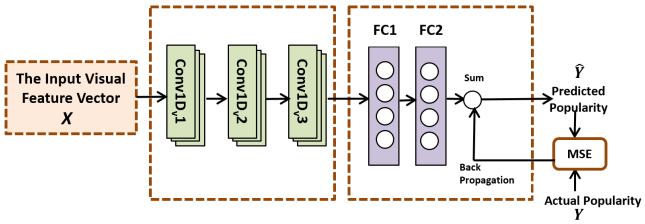


Fig. 4.3. Structure of the VCNN model.

DOI:10.6814/NCCU202100261

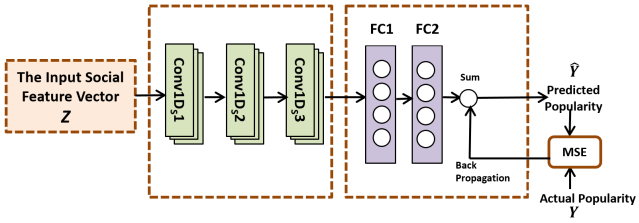


Fig. 4.4. Structure of the SCNN model.

4.4.3 Baseline Models

Considering that the popularity prediction for social media images is a regression problem, only a small number of current machine learning models can be directly used. Therefore, we first compare the proposed VSCNN model with two CNN-based models; namely the visual-only CNN (VCNN) and the social-only CNN (SCNN) model. Then, we compare the proposed VSCNN with the following four conventional regression models: linear regression (LR), SVR, decision tree regression (DTR), and gradient boosting decision tree (GBDT). A brief description of each model is presented in the following subsections.

Visual-only Convolutional Neural Network

As shown in Fig. 4.3, the VCNN model disjoints all social-related parts in the proposed VSCNN (cf. Fig. 4.2 (b)) and retains the remainder. The VCNN is trained using the same procedure as in the VSCNN model, which was presented in Section 4.4.2.

Social-only Convolutional Neural Network

The structure of the SCNN model is shown in Fig. 4.4. It detaches all visual-related components in the proposed VSCNN model (cf. Fig. 4.2 (b)).

DOI:10.6814/NCCU202100261

Linear Regression

LR is a statistical model designed for modeling the relationship between a single dependent variable (output) and a set of independent variables (inputs) by finding a linear regression function that best describes the input variables. To predict the popularity score of an image using this model, a linear relationship between the features of an input image and the popularity score was assumed as follows:

$$y = w_0 + w_1 x_1 + ... + w_n x_n + \varepsilon = \mathbf{x}^{\mathsf{T}} \mathbf{w} + \varepsilon, \tag{4.15}$$

where y indicates the predicted popularity score of the input image, $\mathbf{x}$ denotes the feature vector, $\mathbf{w}$ is the model weight vector, and ε is the error term. The gradient descent algorithm [161] was employed to learn the weight coefficients during the training phase.

Support Vector Regression

SVR [105] is a regression version of a support vector machine [162]. It can construct advanced optimal approximation functions using training data. Given M training samples of popularity feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M\}$, where $\mathbf{x}_i \in \mathbb{R}^d$, and their corresponding popularity score values $\{y_1, y_2, ..., y_M\}$, where $y_i \in \mathbb{R}$, the regression is performed by determining a continuous mapping $f : \mathbb{R}^d \to \mathbb{R}$ that best predicts the set of training samples with the approximation function $y = f(\mathbf{x})$. This is defined as follows:

DOI:10.6814/NCCU202100261

$$y = f(\mathbf{x}) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b, \tag{4.16}$$

where α_i and α_i^* are the Lagrange multipliers associated with each training sample $\mathbf{x}_i$, K denotes the kernel function, and b is the bias term. In this study, the Gaussian radial basis function (RBF) [163] was utilized as the kernel function.

Decision Tree Regression

A decision tree can be used to predict the value of a continuous dependent variable from a set of continuous predictors by constructing a predictive model with a tree-like structure. In this study, the classification and regression tree (CART) algorithm [164] was used to construct a decision tree. Using this algorithm, we constructed a model that can predict the popularity score by learning simple decision rules derived from the data features. For each feature, the CART algorithm splits the data at different points, then selects the part that minimizes the sum of squared errors (SSE) and generates more homogeneous subsets. The splitting process results in a fully grown tree such that the value (popularity score) obtained at each terminal node (leaf node) is the mean of all label values at the node.

Gradient Boosting Decision Trees

GBDT is a machine learning algorithm that recursively constructs an ensemble of weak decision tree models using boosting [107]. It has been proven to be highly efficient in various data mining competitions [165, 166]. The general principle of the GBDT algorithm is the

DOI:10.6814/NCCU202100261

sequential training of a series of simple decision tree estimators [164], where each successive tree attempts to minimize a certain loss function formed by the preceding trees. That is, in each stage, a new regression tree is sequentially added and trained based on the residual error of the previous ensemble model. The GBDT algorithm then updates all the predicted values by adding the predicted values of the new tree. This process is recursively continued until a maximum number of trees have been generated. Thus, the final prediction value of a single instance is the sum of the predictions of all the regression trees.

4.5 Experiments and Results

In this section, we present the experimental setting and discuss the results.

4.5.1 Experimental Setup

Popularity Measurement

Social media websites allow users to interact with posted content in various ways, which results in different social signals that can be utilized to measure the popularity of social content (e.g., images, texts, and videos) on these websites. For instance, on Twitter, popularity can be gauged by the number of re-tweets, whereas the number of likes or comments can be used to measure popularity on Facebook. In this study, we use Flickr as the major image-sharing platform to predict the popularity of social media images. Previous studies have used various metrics to measure the image popularity on Flickr. For example, Khosla *et al.* [38]

DOI:10.6814/NCCU202100261

determined the popularity of an image based on the number of views it received. McParlane *et al.* [41] adopted both view and comment count as the principal metrics.

The dataset used in our experiments complies with Khosla *et al.* [38], and the number of views was adopted as a popularity metric. The log function is applied to manage the large variation in the number of views for various photos from the dataset. Moreover, the images receive views during the time they are online. Thus, a log-normalization approach was used to normalize the effect of the time factor. The score proposed in [38] can be defined as follows:

$$\text{Score}_i = \log_2\left(\frac{p_i}{d_i}\right) + 1, \tag{4.17}$$

where p_i is the popularity metric (the original number of views) of image i, and d_i is the number of days since the image first appeared on Flickr.

Parameter Setting of Baseline Models

All the baseline models were implemented using the scikit-learn machine learning library [167, 168]. In the experiments, the performance of the baseline models was observed to be significantly influenced by several hyper-parameters. Therefore, we identified the values of a few important SVR parameters as follows: C = 3, epsilon = 0.1, gamma = auto, and kernel = RBF. Regarding the DTR model, the best performance was achieved when the max_depth parameter was set to 10. Moreover, we identified several parameters of GBDT: n_estimators = 2000, max_depth = 10, and learning_rate = 0.01. Finally, the remaining parameters are set to their default values in all the models.

DOI:10.6814/NCCU202100261

Dataset

The Social Media Prediction (SMP-T1) dataset presented by ACM Multimedia Grand Challenge in 2017 was used as a real-world dataset to evaluate the performance of the proposed approach [169, 112]. The dataset consists of approximately 432K posts collected from the personal albums of 135 different users on Flickr. Every post in the dataset has a unique picture id along with the associated user id that signifies the user who posted the picture. Additionally, the following image metadata were provided: post date (postdate), number of comments (commentcount), number of tags in the post, whether the photo is tagged by some users or not (haspeople), and character length of the title and image caption (titlelen or deslen). Furthermore, user-centric information, namely the average view count, group count, and average member count, was also provided in the dataset. Each image has a label representing its popularity score (log-normalized views of the image). A few images selected from the dataset are shown in Fig. 4.5. In our experiments, 60% of the images were used for training, 20% for validation, and 20% for testing.

Fig. 4.5. Sample images from the dataset. The popularity of the images is sorted from more popular (left) to less popular (right).

DOI:10.6814/NCCU202100261

Evaluation Metrics

In this study, we used the same following metrics as those in the ACM Multimedia Grand Challenge [169, 112] to assess the prediction accuracy: Spearman's Rho [159], MSE, and mean absolute error (MAE).

- **Spearman's Rho:** Used to calculate the correlation between the predicted popularity scores and the actual scores for the set of tested images. It is defined by the following equation:

$$r_s = 1 - \frac{6 \sum_{j=1}^{n} d_j^2}{n(n^2 - 1)} \tag{4.18}$$

where n is the size of the test sample and d_j is the difference between the two ranks of the actual and predicted popularity values of each image j.

- **MSE:** Usually used to measure the average of the sum of squared prediction errors. Each prediction error represents the difference between the actual value of the data point and the predicted value obtained by the regression model. MSE consists of simple mathematical properties, making it easier to calculate its gradient. In addition, it is often presented as a default metric for most predictive models because it is smoothly differentiable, computationally simple, and hence can be better optimized. A significant limitation of MSE is the fact that it heavily penalizes large prediction errors by squaring them. Because each error in MSE grows quadratically, the outliers in the data significantly contribute to the total error. This indicates that MSE is sensitive to

DOI:10.6814/NCCU202100261

outliers and applies excessive weight on their effects, which leads to an underestimation of the model performance. The drawback of MSE only becomes evident when there are outliers in the data, in which case using MAE is a sufficient alternative. Formally, the MSE is mathematically defined as follows:

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left(\hat{Y}_j - Y_j\right)^2 \tag{4.19}$$

where, in our case, n is the size of the test sample, $\hat{Y}_j$ and Y_j are the predicted and actual popularity scores of j-th image, respectively.

- **MAE:** A simple measure usually used to evaluate the accuracy of a regression model. It measures the average of the absolute values of the individual prediction errors of the model over all samples in the test set. In the MAE metric, each prediction error contributes proportionally to the total amount of errors, indicating that larger errors contribute linearly to the overall error. Because we use the absolute value of the prediction error, the MAE does not indicate underperformance or overperformance of the model, that is, whether the regression model overpredicts or underpredicts the input samples. Thus, it offers a relatively impartial comprehension of how the model performs. By taking the absolute value of the prediction error and not squaring it, the MAE becomes more robust than MSE in managing outliers because it does not heavily penalize the large errors, as done by using MSE. Hence, MAE has its advantages and disadvantages. On one hand, it assists in handling outliers; however, on the other hand, it fails to penalize the large prediction errors. If $\hat{Y}_j$ is the predicted popularity value

DOI:10.6814/NCCU202100261

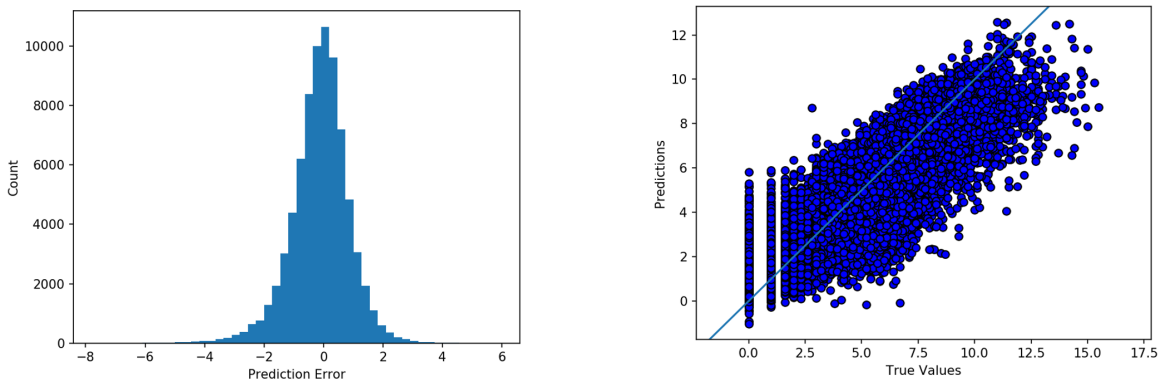
of the j-th sample, and Y_j is the corresponding actual popularity value, then the MAE estimated over a test sample of size n is defined as follows:

$$MAE = \frac{1}{n}\sum_{j=1}^{n} \hat{Y}_j - Y_j \tag{4.20}$$

4.5.2 Results

Using the features extracted for model learning, we trained the proposed VSCNN model to predict the popularity score. In the training stage, we used Adam [170] and the stochastic gradient descent as the learning optimizer to obtain the initialized parameters for VSCNN. The initial learning rate was set to 0.001. In the experiments, the model was run for 50 training epochs over the entire training set. In each epoch, the model was iterated over batches of the training set, where each batch consisted of 20 samples. Furthermore, the following features were added to the training process: 1) The learning rate was reduced by 0.1 every 10 epochs using the learning rate scheduler function, which facilitates learning. 2) The best validation accuracy was saved using the model checkpoint function, which assists in saving the best learning model. The cost function generally converges during the training phase. In the testing stage, the trained VSCNN model was applied to the test samples for evaluation. The evaluation results demonstrated that VSCNN can achieve a Spearman's Rho of 0.9014, an MAE of 0.73, and an MSE of 0.97, which are listed in Tables 4.4 and 4.5, and will be used for comparison with the baseline models.

DOI:10.6814/NCCU202100261



(a) (b)

Fig. 4.6. Quality evaluation of the VSCNN model. (a) Error distribution histogram of the model, and (b) scatterplot of true values (x-axis) versus predicted values (y-axis).

Essential visual analytics were added for the model quality evaluation by computing the error distribution histogram, which presents the distribution of the errors made by the model when predicting the popularity score for each test sample, as shown in Fig. 4.6 (a). A larger number of errors close to zero in the histogram indicates a higher prediction accuracy. Moreover, Fig. 4.6 (b) presents a scatterplot of the actual values on the x-axis versus the predicted values obtained by the model on the y-axis. This scatterplot presents the correlation between the actual and predicted values. If the data appear to be near a straight diagonal line, it indicates a strong correlation. Thus, a perfect regression model would yield a straight diagonal line from the data. From the results shown in Fig. 4.6, there are certain outliers that are not correctly predicted by the VSCNN model. Hence, we analyze these outliers below and explain in detail why our model fails to predict them.

In certain regression problems, the distribution of the target variable may have outliers (e.g., large or small values far from the mean value), which can affect the performance of the

DOI:10.6814/NCCU202100261

predictive model. As shown in Fig. 4.7, the distribution of the target variable (view counts) of the training samples is highly non-uniform in our dataset; therefore, the proposed model attempts to minimize the prediction errors of the largest cluster of view counts of training samples. However, as the number of training samples with extremely high view counts is relatively low, it is more likely that the proposed model cannot correctly predict the high view counts, which will be observed as outliers in the predictive results.

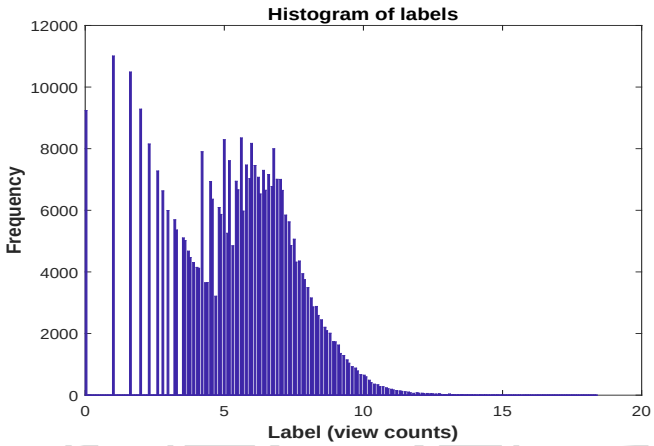
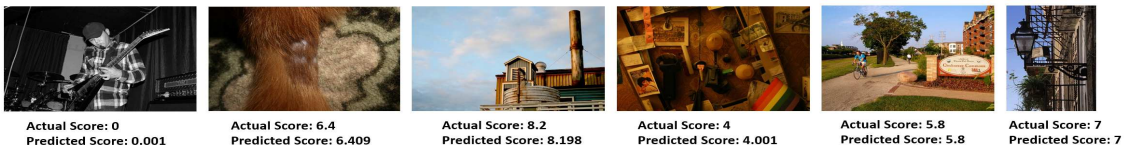


Fig. 4.7. A distribution of the view counts of the training samples.

As shown in Fig. 4.8, a few good and bad predictions were made on images from the test set using our proposed model. The correctly predicted examples are shown in Fig. 4.8 (a); note, our model achieves superior performance with only 0.001 - 0.009 errors relative to the actual scores. For example, the popularity score of the first four images in Fig. 4.8 (a) is correctly predicted with errors of 0.001, 0.009, 0.002, and 0.001, respectively. In addition, the popularity score of the last two images in Fig. 4.8 (a) is perfectly predicted with zero prediction error. On the other hand, a few wrongly predicted examples are shown in Fig. 4.8 (b). For example, the actual popularity score of the first image in this figure is 3, while the

DOI:10.6814/NCCU202100261



(a) Correct examples of popularity prediction

(b) Wrong examples of popularity prediction

Fig. 4.8. Examples of correct and wrong predictions of some images from our dataset using the VSCNN model. The actual popularity score and its corresponding predicted score are displayed below each image.

score obtained by our model is 7.472, resulting in a substantial error of 4.472 in prediction. This disparity is due to the strong indications of some user features for this image, such as average views and member count, which have values of 993.42 and 10,672, respectively, and significantly contribute to the model prediction when integrating all features. Likewise, the last two images in Fig. 4.8 (b) are other badly predicted examples of our proposed model. The actual popularity scores of these two images are observed to be too high. Therefore, it is suggested that our model cannot correctly predict the popularity of these images because the number of training samples with high popularity scores is extremely limited in our dataset, as shown in Fig. 4.7.

Comparison with Baseline Models

First, we train the SCNN model using three different types of social features to explore the influence of each type on predicting popularity. Subsequently, the SCNN model is

DOI:10.6814/NCCU202100261

trained using all the social features as inputs. The prediction results of the SCNN model with different types of input features are summarized in Table 4.4, which presents that the user features perform exceptionally well in predicting the popularity of an image relative to the other two types of social features (i.e., post metadata and time), with a Spearman's Rho of 0.7537, MAE of 1.13, and MSE of 2.17. This indicates that the popularity of an image is closely related to the popularity of the user uploading it; images shared on social media by popular users have a higher chance of obtaining more views. However, not all images posted by popular users are popular. To justify this, we use the popularity score and average view count as popularity metrics for images and users, respectively. As indicated in previous studies [41, 171], the Pareto Principle (or 80-20 rule) was used to select a threshold to differentiate images with high (20%) and low (80%) popularity scores. Likewise, we set a threshold to differentiate users with a high (20%) and low (80%) average view count. Based on these differentiations, the top 20% of the images and users are considered as highly popular (or popular), while the remaining 80% are considered less popular (or common). Accordingly, on average, 69.19% and 16.17% of the images posted by popular and common users, respectively, were determined to be popular. Thus, we conclude that not all images posted by popular users are always popular.

The post metadata are also noteworthy features. The SCNN model using these features achieved values of 0.6590, 1.35, and 2.98 for the Spearman's Rho, MAE, and MSE, respectively. This indicates that image-specific social features, such as tag count, title length, description length, and comment count, also play an important role in predicting popularity, which is expected; an image with significant tags or a longer description/title tends to be

DOI:10.6814/NCCU202100261

Table 4.4 Performance comparison of SCNN, VCNN, and VSCNN models.

Models	Features	Spearman's Rho	MAE	MSE
SCNN	User	0.7537	1.13	2.17
	Post_Metadata	0.6590	1.35	2.98
	Time	0.5317	1.42	3.43
	All_Social	0.8809	0.79	1.13
VCNN	Color	0.3278	1.66	4.46
	Gist	0.2612	1.72	4.67
	LBP	0.3287	1.66	4.45
	Aesthetic	0.2000	1.77	4.91
	Deep	0.4101	1.61	4.13
	All_Visual	0.4168	1.58	4.08
VSCNN	Visual+Social	**0.9014**	**0.73**	**0.97**

more popular because it has a greater chance of showing up in the search results when people use keywords to search for images. Similarly, having more comments on the image suggests that more users interact with the image, which may lead to a greater number of views and thus, increased popularity. Considering the results, time features were also determined to make a significant contribution to popularity prediction, which indicates that the time when an image is posted may influence its popularity. For example, users tend to browse social networking sites at a particular time of the day, such as weekend leisure time, which indicates that images posted during that time are more likely to receive a large number of views and become popular.

Furthermore, while each type of social feature performs sufficiently, the SCNN model achieves the best predictive performance when all the social features are combined, as shown in the fourth row of Table 4.4. This suggests that all the social features proposed are strongly correlated and provide complementary information to each other. Fig. 4.9 (a) presents a diagram of the predicted values obtained by SCNN and the corresponding actual values.

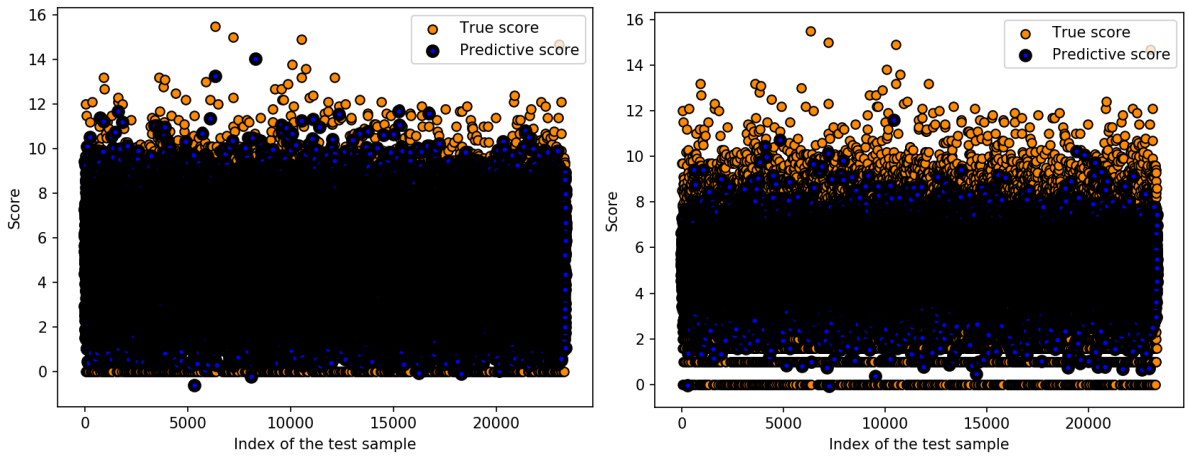
DOI:10.6814/NCCU202100261

Similarly, the VCNN model was trained using each of the individual visual features to analyze their effect on predicting image popularity. We also integrated all the visual features and used them as the input to the model. The evaluation results are listed in Table 4.4. Deep learning features were observed to outperform other visual features. However, it is important to note that the VCNN model achieves the best performance in terms of all the evaluation metrics when all the visual features are combined. In addition, as indicated by the results of the VCNN model, visual features are less effective than social features in terms of image popularity prediction. This finding is consistent with previous studies [38, 172, 43, 173]. Nevertheless, the visual features are useful when there are no post metadata existing, or to address scenarios where no social interactions were recorded prior to publishing the image (e.g., user newly joined social network). This indicates that image content also plays a critical role in popularity prediction, and may complement the social features.

A diagram of the predicted values obtained using VCNN and the corresponding actual values is shown in Fig. 4.9 (b). Finally, the performance of the proposed model is compared with the best performance of both VCNN and SCNN in terms of all the evaluation metrics; the results are listed in Table 4.4. Apparently, VSCNN outperforms VCNN and SCNN, with a relative improvement of 2.33% (SCNN) and 116.27% (VCNN) in terms of Spearman's Rho, and a decrease of 7.59% (SCNN) and 53.80% (VCNN), as well as 14.16% (SCNN) and 76.23% (VCNN) in terms of MAE and MSE, respectively.

Subsequently, the other four baseline models (i.e., LR, SVR, DTR, and GBDT) were trained using each single feature and various combinations thereof to demonstrate the effectiveness of the proposed features in predicting image popularity. The predictions are

DOI:10.6814/NCCU202100261



(a) (b)

Fig. 4.9. Diagrams of the predicted values obtained using the CNN-based baseline models and their corresponding ground truth values. (a) SCNN, and (b) VCNN.

Table 4.5 Performance comparison of LR, SVRR, DTR, GBDT, and VSCNN models.

Features	LR			SVR			DTR			GBDT			VSCNN		
	Spearman's Rho	MAE	MSE	Spearman's Rho	MAE	MSE	Spearman's Rho	MAE	MSE	Spearman's Rho	MAE	MSE	Spearman's Rho	MAE	MSE
Color	0.0856	1.81	5.10	0.2569	1.72	4.74	0.1915	1.76	4.93	0.3381	1.66	4.41			
Gist	0.1337	1.79	5.04	0.3209	1.67	4.55	0.1436	1.80	5.12	0.3176	1.69	4.51			
LBP	0.1546	1.79	5.04	0.3028	1.69	4.63	0.1640	1.78	5.01	0.3126	1.68	4.52			
Aesthetic	0.1221	1.8	5.06	0.1866	1.77	4.97	0.1661	1.79	5.00	0.2040	1.77	4.88			
Deep	0.3701	1.66	4.43	0.4754	1.53	3.88	0.2330	1.76	4.96	0.4403	1.59	4.05			
All_Visual	0.3837	1.65	4.35	0.5018	1.50	3.73	0.2384	1.76	4.95	0.4890	1.53	3.82			
user	0.6449	1.41	3.17	0.7548	1.12	2.18	0.7579	1.12	2.15	0.7580	1.12	2.15			
Post_Metadata	0.5266	1.68	4.41	0.6126	1.42	3.28	0.6682	1.32	2.91	0.6962	1.27	2.72			
Time	0.1337	1.80	5.00	0.2681	1.70	4.65	0.3485	1.61	4.19	0.6285	1.29	2.84			
All_Social	0.7114	1.28	2.72	0.8292	0.94	1.58	0.8049	1.01	1.74	0.8611	0.86	1.30			
Visual+Social	0.7341	1.22	2.44	0.8347	0.93	1.54	0.8200	0.96	1.65	0.8778	0.80	1.15	**0.9014**	**0.73**	**0.97**

shown in Table 4.5, presenting that the user feature yields the best results. This indicates that the characteristics of the person who posts a photo determine its popularity to a significant extent. Furthermore, post metadata and time features were also determined to be sufficient predictors. Additionally, when all social context features are combined and used as inputs,

DOI:10.6814/NCCU202100261

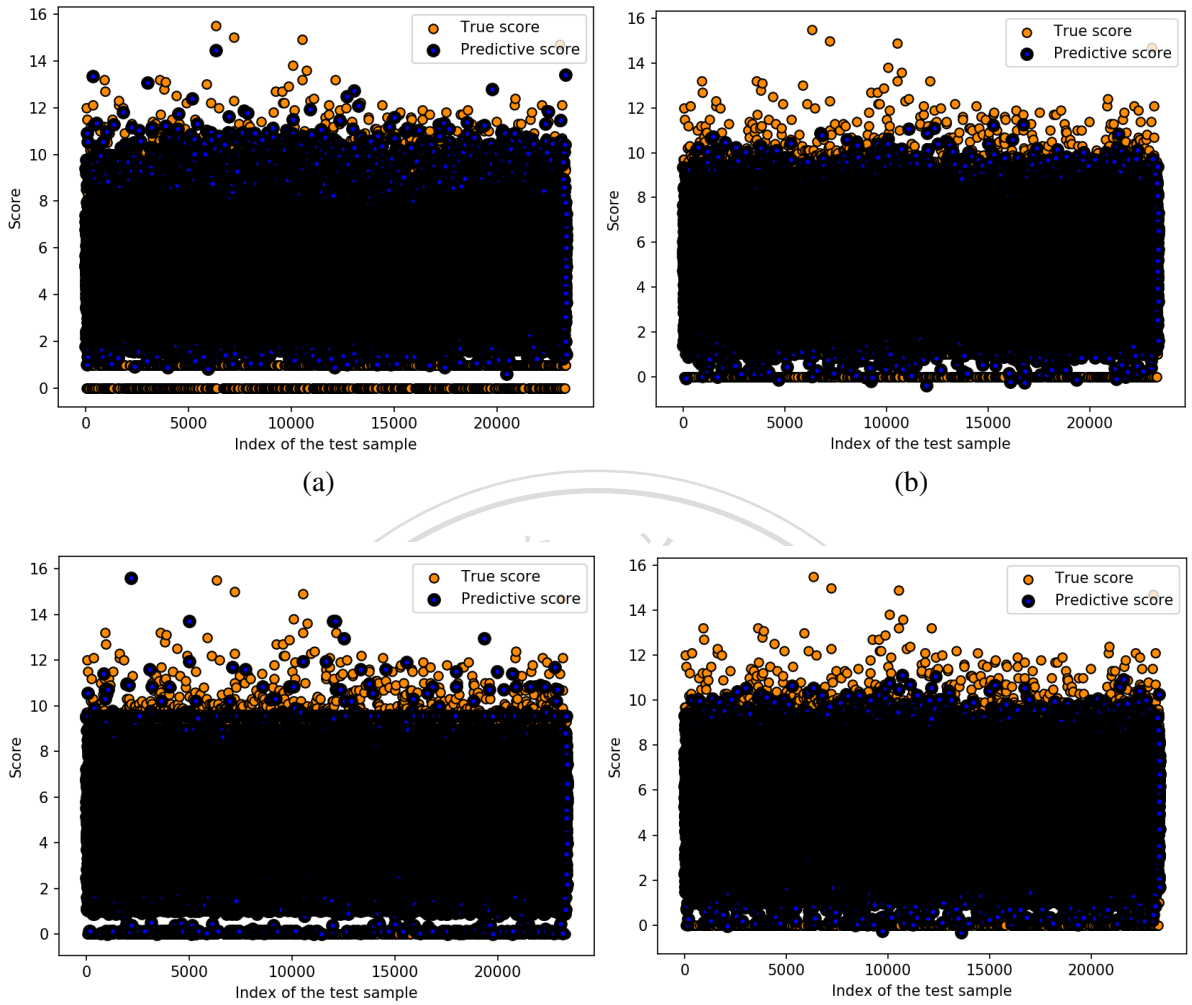


Fig. 4.10. Diagrams of the predicted values obtained using the four machine learning baseline models and their corresponding ground truth values. (a) LR, (b) SVR, (c) DTR, and (d) GBDT.

the performance improves significantly for all the models, and GBDT achieves the best performance in terms of all the evaluation metrics.

The deep learning feature is significant and outperforms other visual features; namely, color, gist, LBP, and aesthetics, although these features perform sufficiently in all models. Nevertheless, the performance of all models is improved when all visual features are

DOI:10.6814/NCCU202100261

combined. Moreover, note that combining visual and social features leads to a significant improvement in the performance of all the models compared to that exhibited using either set of these features independently.

Fig. 4.10 presents diagrams of the predicted values obtained using the four machine learning baseline models and their corresponding ground truth values. As presented in Table 4.5 and shown in Fig. 4.10, GBDT outperformed all other machine learning models, with a relative improvement from 5.16% (SVR) to 19.57% (LR) in terms of Spearman's Rho, and with decreases from 13.98% to 34.43% and from 25.32% to 52.87% in terms of MAE and MSE, respectively. Finally, the performance of the proposed VSCNN model was compared with the best performance obtained by each of the four baseline models (LR, SVR, DTR, and GBDT); the results are shown in Table 4.5. Compared with GBDT, VSCNN improves the prediction performance by approximately 2.69%, 8.75%, and 15.65% in terms of Spearman's Rho, MAE, and MSE, respectively.

Fig. 4.11 presents the best prediction performance for all the models in terms of the three evaluation metrics; the VSCNN outperforms all six baseline models in predicting the popularity of an image. Overall, VSCNN achieved the best prediction performance, with the highest Spearman's Rho (0.9014) and lowest MAE and MSE (0.73 and 0.97, respectively). This suggests that CNNs are more powerful than other machine learning methods in processing heterogeneous information for popularity prediction. Another significant finding is that both social and image content features are essential and complement each other in predicting image popularity on photo-sharing websites.

DOI:10.6814/NCCU202100261

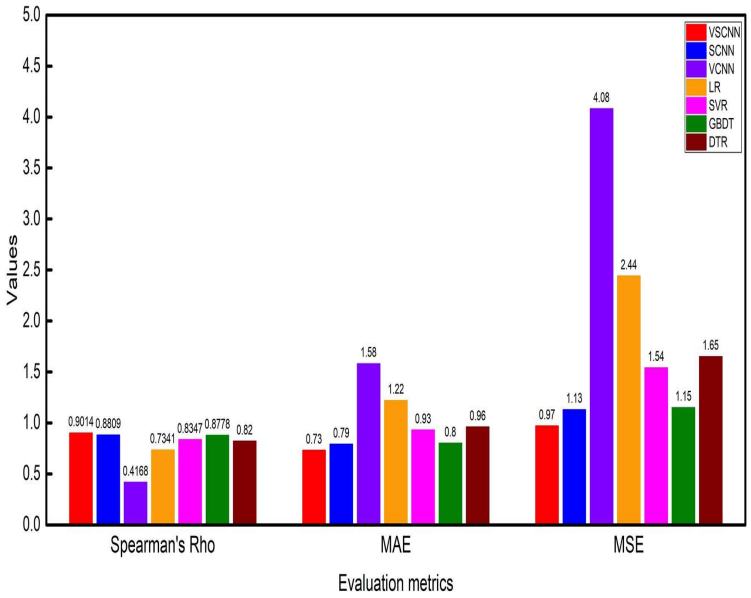


Fig. 4.11. Best prediction performances for all the models in terms of Spearman's Rho, MAE, and MSE metrics.

Comparison with state-of-the-art methods on SMP-T1 dataset

The ACM Multimedia Grand Challenge in 2017 presented a social media prediction task (SMP-T1) as a challenge [169, 112], to predict the popularity of images posted by users on social media. Several teams participated in this challenge, and proposed different models based on the provided SMP-T1 dataset. We compared the performance of the proposed VSCNN model with that of these models, for which the evaluation results are listed in Table 4.6; VSCNN outperforms all the other models. Compared to the best team model (i.e., TaiwanNo.1 SMP-T1), VSCNN improves the prediction performance by approximately 9.02%, 31.62%, and 52.75% in terms of Spearman's Rho, MAE, and MSE, respectively.

In addition to the aforementioned compared models, the multimodal approach presented in [113] integrates significant multimodal information extracted from the same SMPT1

DOI:10.6814/NCCU202100261

Table 4.6 Comparison with the state-of-the-art methods on SMP-T1 dataset.

Methods		Spearman's Rho	MAE	MSE
SMP Challenge Teams [169, 112]	TaiwanNo.1 SMP-T1	0.8268	1.0676	2.0528
	heihei SMP-T1	0.8093	1.1059	2.1767
	NLPR_MMC_Passerby SMP-T1	0.7927	1.1783	2.4973
	BUPTMM SMP-T1	0.7723	1.1733	2.4482
	bluesky SMP-T1	0.7406	1.2475	2.7293
	WePREdictIt SMP-T1	0.5631	1.6278	4.2022
	FirstBlood SMP-T1	0.6456	1.6761	6.3815
	ride_snail_to_race SMP-T1	-0.0405	2.4274	9.2715
	CERTH-ITI-MKLAB SMP-T1	0.3554	3.8178	19.3593
Multimodal approach [113]		0.75	1.12	2.39
VSCNN		**0.9014**	**0.73**	**0.97**

dataset into a CNN model for predicting the popularity of images. Although this approach adopts multimodal features (e.g., image, textual, contextual, and social features) for popularity prediction, it ignores other important features, such as low-level computer vision, aesthetics, and time features. It also adopts an early fusion scheme to merge the features extracted from various modalities into a single large feature vector prior to feeding them into the CNN regression model. Although early fusion can create a joint representation of the input features from multiple modalities, it requires the features to be extremely engineered and preprocessed to be aligned well before the fusion process. It also suffers from the difficulty of representing the time synchronization between multimodal features. Moreover, the increase in the number of modalities makes it difficult to learn the cross-correlation among the overly heterogeneous features. Eventually, a single model is used to make predictions by assuming that the model is well suited for all modalities. However, the architecture of the CNN model used in [113] is not sufficiently powerful to process features from different modalities and then accurately predict the popularity of an image.

DOI:10.6814/NCCU202100261

Unlike [32], our model employs a late fusion scheme in which the features of each modality (i.e., visual and social features) are examined and trained independently using two CNNs with a highly designed architecture. The obtained results are then fused using a merged layer into another network for further processing and obtaining the final prediction. The fusion process in our model becomes easy to execute and does not suffer from the data representation problem that the early fusion scheme has because the semantic vectors resulting from the two CNN models usually have the same form of data. In addition, late fusion allows the usage of the most suitable model for analyzing each modality and learning its features, providing more flexibility. Furthermore, the robust interpretation of incomplete and inconsistent multimodal input becomes more reliable at later stages because more semantic knowledge becomes available from various sources. Owing to these advantages, the late fusion scheme is extensively used in multimodal systems [38, 114, 172]. To confirm the efficiency of our model, we also compared it to the multimodal approach proposed in [113]; the prediction results are summarized in Table 4.6. Apparently, VSCNN outperforms the multimodal approach, with a relative improvement of 20.19% in terms of Spearman's Rho, and a decrease of 34.82% and 59.41% in terms of MAE, and MSE, respectively.

Late and Early Fusion Schemes for the VSCNN model

The framework proposed in this study adopts the late fusion scheme; that is, we first employ two convolutional neural networks to process visual features and the corresponding social context information individually. Then, the outputs of these two networks are merged into another network that fully connects all the information into a final layer of the deep

DOI:10.6814/NCCU202100261

architecture. We also tested the early fusion scheme by integrating visual and social features at the input of the convolutional layers. The early fusion scheme, denoted by VSCNN-EF, replaces the visual and social networks in Fig. 4.2 with a unified CNN whose inputs comprise of fused visual-social features obtained by concatenating the visual and social features of a given image into a final feature vector of 4,744 dimensions (4,710 visual and 34 social). Then, PCA [160] is applied to reduce the dimensionality of this vector from 4,744 to 20 and to select only the most prevalent features. The numbers of parameters of VSCNN and VSCNN-EF are of the same order. These schemes are optimized and tested, followed by comparing the prediction performance in terms of the three performance metrics; the results are listed in Table 4.7. Apparently, VSCNN consistently outperforms VSCNN-EF, suggesting that the proposed late fusion scheme, which initially processes visual and social information independently and merges them later, is better than an early fusion scheme, which incorporates the heterogeneous data at the beginning.

Table 4.7 Performance comparison of VSCNN and VSCNN-EF models.

Models	Spearman's Rho	MAE	MSE
VSCNN	**0.9014**	**0.73**	**0.97**
VSCNN-EF	0.8898	0.76	1.07

4.6 Summary

Recently, deriving an effective computational model to characterize human behavior or predict decision making has become an emergent topic. In this chapter, we developed a multimodal deep learning framework for predicting the popularity of images on social

DOI:10.6814/NCCU202100261

media. First, we analyzed and extracted different types of image visual content features and social context information that significantly affect image popularity. Then, we proposed a novel CNN-based visual-social computational model for image popularity prediction, called VSCNN. This model uses individual networks to process input data with different modalities (i.e., visual and social features), and the outputs from these networks are then integrated into a fusion network to learn joint multimodal features and estimate the popularity score. We trained the proposed model in an end-to-end manner. The experimental results on the provided dataset demonstrate the effectiveness of the proposed model in predicting image popularity. Further experiments demonstrated that VSCNN achieved a notably superior prediction performance. Specifically, it outperformed four traditional machine learning schemes, two CNN-based models, and other state-of-the-art methods, in terms of three standard evaluation metrics (i.e., Spearman's Rho, MAE, and MSE). This emphasizes the effectiveness of the proposed model in combining visual and social information to predict the popularity of an image.

DOI:10.6814/NCCU202100261

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This dissertation has attempted to explore and understand human social interaction in both the real and online worlds. In real life, the human face is one of the most powerful tools in social interaction, from which viewers can rapidly and easily make a number of inferences—about age, gender, race, emotional state, physical health, etc. Of these inferences, age is one of the most important face traits used in interpersonal communication and interaction in our social life, as our interaction and responding way are radically dependent on the relative age of our interlocutors. Humans usually rely on the facial appearance trait for age estimation in everyday life, as the appearance of human faces displays notable changes with the aging progress. However, the human estimation of facial age is usually not as precise as other types of facial traits. This implies that there is a high desire for developing automatic facial

DOI:10.6814/NCCU202100261

age estimation methods that are comparable or even superior to the human ability in age estimation. To this end, in Chapter 3, we propose a comparative deep learning framework for facial age estimation, named CRCNN. The basic idea behind this framework is that the input face is not directly used to estimate the age, but it first compares with a set of labeled baseline samples to create a set of hints (comparative relations, i.e., the input face is younger or older than each of the baseline sample). Then, the estimation stage aggregates all the set of hints and then votes on the number of hints for each label to estimate the person's age. In comparison to the traditional approach, we notice that this comparative approach has several advantages: (1) the age estimation task is divided into several comparative stages, which is much simpler than directly guessing the exact age of an observed face; (2) by increasing the number of baseline samples, more side information can be provided to robust the age estimation task; and (3) by leveraging numerous baseline samples, few incorrect comparisons will significantly not affect the accuracy of the age estimation. In addition, the recent method of auxiliary coordinates (MAC) is integrated for training the comparative stage. This method presents a set of variables to break the objective function dependency, making the problem much better conditioned without nesting, affording an efficient and distributed optimization. The experimental results show that our proposed CRCNN framework significantly outperforms other state-of-the-art methods.

On the other hand, it is noteworthy that the basic function of social media websites is to facilitate and increase social interaction on the internet. Although current online social networks have varied functionalities and objectives, they share similar interaction modes: public interaction mode and private interaction mode. The former refers to social interaction

DOI:10.6814/NCCU202100261

with the published content on these websites (e.g., by commenting, liking, or viewing), which is public, while the latter refers to direct communication between relationship parties, such as through Facebook inbox and Twitter message. Overall, we all concur that social media websites have entirely changed our way of being and our interactions with others. The explosive growth in the number of users, volume of activities, and forms of interaction on social networks often make some of the content published on these networks more popular than others. For example, when we consider images on social photo-sharing websites, we notice a variance in their number of views or "social popularity", irrespective of their visual appeal. This variation is observed even among images posted by the same user, or images from the same category. This poses the following two questions: What are the factors that make an image popular on social media? Can we design an efficient predictive model that can accurately predict the number of views an image will obtain even before it is uploaded? To address these questions, in Chapter 4, we first analyze a real-world dataset collected from Flickr to investigate two critical features that may influence an image's popularity, namely visual content and social context. Then, to demonstrate the effectiveness of the proposed features in predicting image popularity, we propose a multimodal deep learning prediction model, called VSCNN. In the proposed VSCNN model, two individual CNNs are adopted to learn discriminative representations of the visual and social features independently and then efficiently merge them into a unified network for popularity prediction. The experimental results show the predictive power of the proposed visual content and social context features towards image popularity. Moreover, combining multimodal features boosts the performance of the prediction model. Consequently, our proposed method of using a multimodal approach

DOI:10.6814/NCCU202100261

to predict image popularity on photo-sharing websites is more powerful than a single modal approach. Finally, the simulation results also show that the proposed VSCNN model achieved a notably superior prediction performance as compared to the state-of-the-art techniques.

5.2 Future Work

With regard to the CRCNN framework, the experimental results in this dissertation have confirmed that the side information obtained from the baseline samples in the comparative stage alongside the input face information can accelerate the overall performance of the framework and boost the age estimation task. Therefore, one of our future works is to improve the baseline selection, because obtaining an effective baseline is crucial in our comparative approach. In addition, as aging procedures are considerably different from person-to-person, especially from different social groups, we plan to build a "baseline bank" (constituted by a set of baselines that corresponding to a computed group of social consistency) instead of using a single and global baseline. Further research on CRCNN in these directions will be considered attractive future work. In our future research, we will also aim to modify the existing CNNs in our framework by considering other pre-trained CNNs, such as VGG-Face CNN [174] and AlexNet CNN [111] to improve estimates of age.

With regard to the VSCNN model, the results explicitly demonstrated that both image content and social context features are essential and complement each other in predicting image popularity on photo-sharing social networks. However, in addition to visual content and social context information, the impact of the textual information associated with the

DOI:10.6814/NCCU202100261

image cannot be neglected when predicting an image's popularity because this information has a primary role in increasing the accessibility of the image when people search for images using keywords. Based on this observation, in our future work, we plan to use a generative model as suggested in [175] to automatically generate natural sentences describing the content and title for each image in the SMP-T1 dataset, and use an image annotation model as proposed in [176] to create a set of keywords (hashtags) that are related to the content of the image. Then, the obtained textual information can be incorporated into our model to explore its effect on image popularity. Furthermore, we will also aim to optimize the parameters and overall structures of the CNNs used in the proposed model to improve the prediction performance. In our future research, we will extend our work by considering not only internal but also external factors that may affect image popularity, such as real-world events. Meanwhile, we will investigate the influence of various aspects on image popularity based on geographical location and cultural background.

DOI:10.6814/NCCU202100261

References

[1] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[2] Maja Pantic, Anton Nijholt, Alex Pentland, and Thomas S Huanag. Human-centred intelligent human? computer interaction (HCI^2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008.

[3] Ahmed Elgammal. Human-centered multimedia: representations and challenges. In *Proceedings of the 1st ACM international workshop on Human-centered multimedia*, pages 11–18. ACM, 2006.

[4] Jim Gemmell, Kentaro Toyama, C Lawrence Zitnick, Thomas Kang, and Steven Seitz. Gaze awareness for video-conferencing: A software approach. *IEEE MultiMedia*, 7(4):26–35, 2000.

[5] Michael Hecht, Joseph De Vito, and Laura Guerrero. Perspectives on nonverbal communication: codes, functions, and contexts. *The Nonverbal Communication Reader*, pages 3–18, 1999.

[6] Dacher Keltner, Paul Ekman, Gian C Gonzaga, and Jennifer Beer. Facial expression of emotion. 2003.

[7] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[8] Albert Mehrabian and Susan R Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology*, 31(3):248, 1967.

DOI:10.6814/NCCU202100261

[9] Jon E Grahe and Frank J Bernieri. The importance of nonverbal cues in judging rapport. *Journal of Nonverbal behavior*, 23(4):253–269, 1999.

[10] Sen Pei, Lev Muchnik, José S Andrade Jr, Zhiming Zheng, and Hernán A Makse. Searching for superspreaders of information in real-world social media. *Scientific reports*, 4:5547, 2014.

[11] Elizabeth Dubois and Devin Gaffney. The multiple facets of influence: Identifying political influentials and opinion leaders on twitter. *American behavioral scientist*, 58(10):1260–1277, 2014.

[12] Ioannis Arapakis, Mounia Lalmas, B Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M Jose. User engagement in online n ews: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology*, 65(10):1988–2005, 2014.

[13] Anjana Susarla, Jeong-Ha Oh, and Yong Tan. Social networks and the diffusion of user-generated content: Evidence from youtube. *Information Systems Research*, 23(1):23–41, 2012.

[14] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 965–974, 2014.

[15] Malcolm Gladwell. Blink: The power of thinking without thinking. 2006.

[16] Mary Lee Hummert, Jaye L Shaner, Teri A Garstka, and Clark Henry. Communication with older adults: The influence of age stereotypes, context, and communicator age. *Human Communication Research*, 25(1):124–151, 1998.

[17] Matthew G Rhodes. Age estimation of faces: A review. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 23(1):1–12, 2009.

[18] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008.

DOI:10.6814/NCCU202100261

[19] Guodong Guo, Yun Fu, Thomas S Huang, and Charles R Dyer. Locally adjusted robust regression for human age estimation. In *2008 IEEE Workshop on Applications of Computer Vision*, pages 1–6. IEEE, 2008.

[20] DS Berry, L Zebrowitz-MeArthur, and TR Alley. Social and applied aspects of perceiving faces. 1988.

[21] Kestutis Sveikata, Irena Balciuniene, Janina Tutkuviene, et al. Factors influencing face aging. literature review. *Stomatologija*, 13(4):113–116, 2011.

[22] Harold Smulyan, Roland G Asmar, Annie Rudnicki, Gerard M London, and Michel E Safar. Comparative effects of aging in men and women on the properties of the arterial tree. *Journal of the American College of Cardiology*, 37(5):1374–1380, 2001.

[23] Emma C Paes, Hans JLJM Teepen, Willemijn A Koop, and Moshe Kon. Perioral wrinkles: histologic differences between men and women. *Aesthetic Surgery Journal*, 29(6):467–472, 2009.

[24] Yun Fu, Ye Xu, and Thomas S Huang. Estimating human age by manifold analysis of face pictures and regression on aging features. In *2007 IEEE International Conference on Multimedia and Expo*, pages 1383–1386. IEEE, 2007.

[25] David A Gunn, Helle Rexbye, Christopher EM Griffiths, Peter G Murray, Amelia Fereday, Sharon D Catt, Cyrena C Tomlin, Barbara H Strongitharm, Dave I Perrett, Michael Catt, et al. Why some women look young for their age. *PloS one*, 4(12):e8021, 2009.

[26] Kai Li, Junliang Xing, Weiming Hu, and Stephen J Maybank. D2c: Deep cumulatively and comparatively learning for human age estimation. *Pattern Recognition*, 66:95–105, 2017.

[27] Bingbing Ni, Zheng Song, and Shuicheng Yan. Web image mining towards universal age estimator. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 85–94, 2009.

[28] Young H Kwon and Niels da Vitoria Lobo. Age classification from facial images. *Computer vision and image understanding*, 74(1):1–21, 1999.

DOI:10.6814/NCCU202100261

[29] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on pattern Analysis and machine Intelligence*, 24(4):442–455, 2002.

[30] Ranjan Jana, Debaleena Datta, and Rituparna Saha. Age estimation from face image using wrinkle features. *Procedia Computer Science*, 46:1754–1761, 2015.

[31] Dong Yi, Zhen Lei, and Stan Z Li. Age estimation by multi-scale convolutional network. In *Asian conference on computer vision*, pages 144–158. Springer, 2014.

[32] Xiaolong Wang, Rui Guo, and Chandra Kambhamettu. Deeply-learned feature for age estimation. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 534–541. IEEE, 2015.

[33] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016.

[34] Shixing Chen, Caojin Zhang, and Ming Dong. Deep age estimation: From classification to ranking. *IEEE Transactions on Multimedia*, 20(8):2209–2222, 2017.

[35] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

[36] Yoshua Bengio. Deep learning of representations: Looking forward. *Dediu, A.-H., Martın-Vide, C., Mitkov, R., Truthe, B. (eds.) SLSP. LNCS*, 7978:1–37, 2013.

[37] John B. Carroll. Human cognitive abilities: A survey of factor-analytic studies. *New York : Cambridge University Press*, 1993.

[38] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd international conference on World wide web*, pages 867–876. ACM, 2014.

[39] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. Image popularity prediction in social media using sentiment and context features. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 907–910. ACM, 2015.

DOI:10.6814/NCCU202100261

[40] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. Latent factors of visual popularity prediction. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 195–202. ACM, 2015.

[41] Philip J McParlane, Yashar Moshfeghi, and Joemon M Jose. Nobody comes here anymore, it's too crowded; predicting image popularity on flickr. In *Proceedings of International Conference on Multimedia Retrieval*, page 385. ACM, 2014.

[42] Luam Catao Totti, Felipe Almeida Costa, Sandra Avila, Eduardo Valle, Wagner Meira Jr, and Virgilio Almeida. The impact of visual attributes on online image diffusion. In *Proceedings of the 2014 ACM conference on Web science*, pages 42–51. ACM, 2014.

[43] Wen Wang and Wei Zhang. Combining multiple features for image popularity prediction in social media. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1901–1905. ACM, 2017.

[44] Zheng Song, Bingbing Ni, Dong Guo, Terence Sim, and Shuicheng Yan. Learning universal multi-view age estimator using video context. In *2011 International Conference on Computer Vision*, pages 241–248. IEEE, 2011.

[45] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Trans. System, Man, and Cybernetics*, 34(1):621–628, 2004.

[46] Yun Fu, Guodong Guo, and Thomas S Huang. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1955–1976, 2010.

[47] Xin Geng, Zhi-Hua Zhou, Yu Zhang, Gang Li, and Honghua Dai. Learning from facial aging patterns for automatic age estimation. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 307–316, 2006.

[48] Feng Gao and Haizhou Ai. Face age classification on consumer images with gabor feature and fuzzy lda method. In *International Conference on Biometrics*, pages 132–141. Springer, 2009.

[49] ElectronicCustomerRelationshipManagement(ECRM),https://en.wikipedia.org/wiki/ECRM, 2020.

DOI:10.6814/NCCU202100261

[50] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Soft biometric traits for personal recognition systems. In *International conference on biometric authentication*, pages 731–738. Springer, 2004.

[51] Eric Patterson, Amrutha Sethuram, Midori Albert, Karl Ricanek, and Michael King. Aspects of age variation in facial morphology affecting biometrics. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6. IEEE, 2007.

[52] K Ricanek Jr, E Boone, and E Patterson. Craniofacial aging impacts on the eigenface face biometric. *Comput. Sci*, 1(3), 2006.

[53] Karl Ricanek and Edward Boone. The effect of normal adult aging on standard pca face recognition accuracy rates. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2018–2023. IEEE, 2005.

[54] Narayanan Ramanathan and Rama Chellappa. Face verification across age progression. *IEEE transactions on image processing*, 15(11):3349–3361, 2006.

[55] Junyan Wang, Yan Shang, Guangda Su, and Xinggang Lin. Age simulation for face recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 913–916. IEEE, 2006.

[56] Alexandru Tatar, Marcelo Dias De Amorim, Serge Fdida, and Panayotis Antoniadis. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1):8, 2014.

[57] Nancy J Gnana Amala and K Kumar. Content popularity prediction methods-a survey. In *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, pages 749–753. IEEE, 2018.

[58] Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang, and Sandeep Pandey. Click-through prediction for advertising in twitter timeline. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1959–1968. ACM, 2015.

[59] Flavio Figueiredo, Jussara M Almeida, Marcos André Gonçalves, and Fabrício Benevenuto. On the dynamics of social media popularity: A youtube case study. *ACM Transactions on Internet Technology (TOIT)*, 14(4):24, 2014.

DOI:10.6814/NCCU202100261

[60] Chun-Che Wu, Tao Mei, Winston H Hsu, and Yong Rui. Learning to personalize trending image search suggestion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 727–736. ACM, 2014.

[61] Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P Gummadi. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*, pages 13–18, 2008.

[62] Eldar Sadikov, Montserrat Medina, Jure Leskovec, and Hector Garcia-Molina. Correcting for missing data in information cascades. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 55–64, 2011.

[63] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936, 2014.

[64] Sung Eun Choi, Youn Joo Lee, Sung Joo Lee, Kang Ryoung Park, and Jaihie Kim. A comparative study of local feature extraction for age estimation. In *2010 11th International Conference on Control Automation Robotics & Vision*, pages 1280–1284. IEEE, 2010.

[65] Jhony K Pontes, Alceu S Britto, Clinton Fookes, and Alessandro L Koerich. A flexible hierarchical approach for facial age estimation based on multiple features. *Pattern Recognition*, 54:34–51, 2016.

[66] Sung Eun Choi, Youn Joo Lee, Sung Joo Lee, Kang Ryoung Park, and Jaihie Kim. Age estimation using a hierarchical classifier based on global and local facial features. *Pattern recognition*, 44(6):1262–1281, 2011.

[67] Ivan Huerta, Carles Fernández, and Andrea Prati. Facial age estimation through the fusion of texture and local appearance descriptors. In *European conference on computer vision*, pages 667–681. Springer, 2014.

[68] Guodong Guo and Guowang Mu. A framework for joint estimation of age, gender and ethnicity on a large database. *Image and Vision Computing*, 2014.

[69] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S Huang. Human age estimation using bio-inspired features. In *2009 IEEE conference on computer vision and pattern recognition*, pages 112–119. IEEE, 2009.

DOI:10.6814/NCCU202100261

[70] X. Geng, W.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(12):2234–2240, 2007.

[71] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. In *European conference on computer vision*, pages 484–498. Springer, 1998.

[72] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[73] A Gunay and Vasif V Nabiyev. Facial age estimation based on decision level fusion of amm, lbp and gabor features. *Int. J. Adv. Comput. Sci. Appl*, 6:19–26, 2015.

[74] Jinli Suo, Tianfu Wu, Songchun Zhu, Shiguang Shan, Xilin Chen, and Wen Gao. Design sparse features for age estimation using hierarchical face model. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008.

[75] Jinli Suo, Feng Min, Songchun Zhu, Shiguang Shan, and Xilin Chen. A multi-resolution dynamic model for face aging simulation. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[76] Renliang Weng, Jiwen Lu, Gao Yang, and Yap-Peng Tan. Multi-feature ordinal ranking for facial age estimation. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[77] Carles Fernández, Ivan Huerta, and Andrea Prati. A comparative evaluation of regression learning algorithms for facial age estimation. In *International Workshop on Face and Facial Expression Recognition from Real World Videos*, pages 133–144. Springer, 2014.

[78] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[79] Rajeev Ranjan, Sabrina Zhou, Jun Cheng Chen, Amit Kumar, Azadeh Alavi, Vishal M Patel, and Rama Chellappa. Unconstrained age estimation with deep convolutional neural networks. In *proceedings of the ieee international conference on computer vision workshops*, pages 109–117, 2015.

DOI:10.6814/NCCU202100261

[80] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[81] Kazuya Ueki, Teruhide Hayashida, and Tetsunori Kobayashi. Subspace-based age-group classification using facial images under various lighting conditions. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 6–pp. IEEE, 2006.

[82] Ying Zheng, Hongxun Yao, Yanhao Zhang, and Pengfei Xu. Age classification based on back-propagation network. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 319–322, 2013.

[83] Yun Fu and Thomas S Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4):578–584, 2008.

[84] Khoa Luu, Karl Ricanek, Tien D Bui, and Ching Y Suen. Age estimation using active appearance models and support vector machine regression. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–5. IEEE, 2009.

[85] Peter N Belhumeur, Joao P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *European conference on computer vision*, pages 43–58. Springer, 1996.

[86] Hu Han and Anil K Jain. Age, gender and race estimation from unconstrained face images. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5)*, 87:27, 2014.

[87] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.

[88] Shuicheng Yan, Huan Wang, Xiaoou Tang, and Thomas S Huang. Learning auto-structured regressor from uncertain nonnegative labels. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007.

[89] Hironori Takimoto, Yasue Mitsukura, Minoru Fukumi, and Norio Akamatsu. Robust gender and age estimation under varying facial pose. *Electronics and Communications in Japan*, 91(7), 2008.

[90] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(10), 2013.

DOI:10.6814/NCCU202100261

[91] Chenjing Yan, Congyan Lang, Tao Wang, Xuetao Du, and Chen Zhang. Age estimation based on convolutional neural network. *Advances in Multimedia Information Processing*, 8879:211–220, 2014.

[92] Furkan Gurpinar, Heysem Kaya, Hamdi Dibeklioglu, and Ali Salah. Kernel elm and cnn based facial age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 80–86, 2016.

[93] Zengwei Huo, Xu Yang, Chao Xing, Ying Zhou, Peng Hou, Jiaqi Lv, and Xin Geng. Deep age distribution learning for apparent age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 17–24, 2016.

[94] Bartłomiej Hebda and Tomasz Kryjak. A compact deep convolutional neural network architecture for video based age and gender estimation. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 787–790. IEEE, 2016.

[95] Jun-Cheng Chen, Amit Kumar, Rajeev Ranjan, Vishal M Patel, Azadeh Alavi, and Rama Chellappa. A cascaded convolutional neural network for age estimation of unconstrained faces. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2016.

[96] T. Joachims. Optimizing search engines using clickthrough data. *International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[97] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, pages 933–969, 2003.

[98] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. *International Conference on Machine Learning*, pages 89–96, 2005.

[99] Samah Aloufi, Shiai Zhu, and Abdulmotaleb El Saddik. On the prediction of flickr image popularity by analyzing heterogeneous social sensory data. *Sensors*, 17(3):631, 2017.

DOI:10.6814/NCCU202100261

[100] Liangjie Hong, Ovidiu Dan, and Brian D Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58. ACM, 2011.

[101] Ethem F Can, Hüseyin Oktay, and R Manmatha. Predicting retweet count using visual cues. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1481–1484. ACM, 2013.

[102] Roja Bandari, Sitaram Asur, and Bernardo A Huberman. The pulse of news in social media: Forecasting popularity. *ICWSM*, 12:26–33, 2012.

[103] Bo Wu, Tao Mei, Wen-Huang Cheng, Yongdong Zhang, et al. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *AAAI*, pages 272–278, 2016.

[104] Khaled Almgren, Jeongkyu Lee, et al. Predicting the future popularity of images on social networks. In *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*, page 15. ACM, 2016.

[105] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

[106] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[107] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[108] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[109] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[110] Fatma S Abousaleh, Tekoing Lim, Wen-Huang Cheng, Neng-Hao Yu, M Anwar Hossain, and Mohammed F Alhamid. A novel comparative deep learning framework for

DOI:10.6814/NCCU202100261

facial age estimation. *EURASIP Journal on Image and Video Processing*, 2016(1):47, 2016.

[111] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Conference on Neural Information Processing Systems*, 2012.

[112] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, and Tao Mei. Sequential prediction of social media popularity with deep temporal context networks. *arXiv preprint arXiv:1712.04443*, 2017.

[113] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. A multimodal approach to predict social media popularity. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 190–195. IEEE, 2018.

[114] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.

[115] Microsoft Corp. *How-Old.net*, 2015.

[116] Tsung-Hung Tsai, Wei-Cih Jhou, Wen-Huang Cheng, Min-Chun Hu, I-Chao Shen, Tekoing Lim, Kai-Lung Hua, Ahmed Ghoneim, M. Anwar Hossain, and Shintami C. Hidayati. Photo sundial: Estimating the time of capture in consumer photos. *Neurocomputing*, 177:529–542, 2016.

[117] Chuang-Wen You, Yi-Ling Chen, and Wen-Huang Cheng. Socialcrc: Enabling socially-consensual rendezvous coordination by mobile phones. *Pervasive and Mobile Computing*, 25:67–87, 2016.

[118] Wen-Huang Cheng, Chia-Wei Wang, and Ja-Ling Wu. Video adaptation for small display based on content recomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(1):43–58, 2007.

[119] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, and Tao Mei. Time matters: Multi-scale temporalization of social media popularity. In *Proceedings of the ACM International Conference on Multimedia*, 2016.

DOI:10.6814/NCCU202100261

[120] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

[121] Tekoing Lim, Kai-Lung Hua, Hong-Cyuan Wang, Kai-Wen Zhao, Min-Chun Hu, and Wen-Huang Cheng. Vrank: Voting system on ranking model for human age estimation. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, 2015.

[122] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[123] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

[124] Miguel A. Carreira-Perpinan and Weiran Wang. Distributed optimization of deeply nested systems. *International Conference on Artificial Intelligence and Statistics*, 33:10–19, 2014.

[125] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint*, arXiv:1408.5093, 2014.

[126] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. *IEEE Conference on Computer Vision and Pattern Recognition*, 1:539–546, 2005.

[127] J. Nocedal and S. J. Wright. Numerical optimization. *Springer Series in Operations Research and Financial Engineering*, 2006.

[128] Juha Ylioinas, Abdenour Hadid, Xiaopeng Hong, and Matti Pietikäinen. Age estimation using local binary pattern kernel density estimate. *International Conference on Image Analysis and Processing*, 8156:141–150, 2013.

[129] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. 2011.

DOI:10.6814/NCCU202100261

[130] K. Ricanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. *International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006.

[131] A. C. Gallagher and T. Chen. Using group prior to identify people in consumer images. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[132] Jordi Sanchez-Riera, Kai-Lung Hua, Yuan-Sheng Hsiao, Tekoing Lim, Shintami C. Hidayati, and Wen-Huang Cheng. A comparative study of data fusion for rgb-d based visual recognition. *Pattern Recognition Letters*, 73:1–6, 2016.

[133] Mu Li, Tong Zhang, Yuqiang Chen, and Alex Smola. Efficient mini-batch training for stochastic optimization. *International Conference on Knowledge Discovery and Data Mining*, 2014.

[134] E. A. Patrick and F. P. Fischer. A generalized k-nearest neighbor rule. *Information and Control*, 16(2):128–152, 1970.

[135] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by backpropagating errors. *Nature*, 323(9):318–362, 1986.

[136] J. R. Quinlan. C4.5: Programs for machine learning. *Morgan Kaufmann*, 1993.

[137] V. Vapnik. Statistical learning theory. *John Wiley and Sons*, 1998.

[138] R. Jang. Anfis: Adaptive network based fuzzy inference system. *IEEE Transaction on System, Man and Cybernetics*, 23(3):665–684, 1993.

[139] Fatma S Abousaleh, Wen-Huang Cheng, Neng-Hao Yu, and Yu Tsao. Multimodal deep learning framework for image popularity prediction on social media. *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[140] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[141] Xiang Niu, Lusong Li, Tao Mei, Jialie Shen, and Ke Xu. Predicting image popularity in an incomplete social media community by a weighted bi-partite graph. In *2012 IEEE International Conference on Multimedia and Expo*, pages 735–740. IEEE, 2012.

[142] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. Rt to win! predicting message propagation in twitter. *ICWSM*, 11:586–589, 2011.

DOI:10.6814/NCCU202100261

[143] Amandianeze O Nwana, Salman Avestimehr, and Tsuhan Chen. A latent social approach to youtube popularity prediction. In *Global Communications Conference (GLOBECOM), 2013 IEEE*, pages 3138–3144. IEEE, 2013.

[144] Henrique Pinto, Jussara M Almeida, and Marcos A Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 365–374. ACM, 2013.

[145] David A Shamma, Jude Yew, Lyndon Kennedy, and Elizabeth F Churchill. Viral actions: Predicting video view counts using synchronous sharing behaviors. In *ICWSM*, 2011.

[146] Niyati Aggrawal, Archit Ahluwalia, Prashi Khurana, and Anuja Arora. Brand analysis framework for online marketing: ranking web pages and analyzing popularity of brands on social media. *Social Network Analysis and Mining*, 7(1):21, 2017.

[147] Marcos André Gonçalves, Jussara M Almeida, Luiz GP dos Santos, Alberto HF Laender, and Virgílio Almeida. On popularity in the blogosphere. *IEEE Internet Computing*, 14(3):42–49, 2010.

[148] Aboul-Ella Hassanien and Ajith Abraham. *Computational Intelligence in Multimedia Processing: Recent Advances*, volume 96. Springer, 2008.

[149] Marko Heikkila and Matti Pietikainen. A texture-based method for modeling the background and detecting moving objects. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):657–662, 2006.

[150] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[151] Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):765–781, 2011.

[152] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

DOI:10.6814/NCCU202100261

[153] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, pages 386–399. Springer, 2008.

[154] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 419–426. IEEE, 2006.

[155] Xiaoou Tang, Wei Luo, and Xiaogang Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8):1930–1943, 2013.

[156] Xiaoqiao Chen, Qingyi Zhang, Manhui Lin, Guangyi Yang, and Chu He. No-reference color image quality assessment: from entropy to perceptual quality. *EURASIP Journal on Image and Video Processing*, 2019(1):77, 2019.

[157] Congcong Li, Alexander C Loui, and Tsuhan Chen. Towards aesthetics: A photo quality assessment and photo selection system. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 827–830. ACM, 2010.

[158] https://keras.io/.

[159] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[160] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[161] JVN Lakshmi. Stochastic gradient descent using linear regression with python. *International Journal of Advanced Engineering Research and Applications*, 2(8):519–525, 2016.

[162] Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[163] Martin Hofmann. Support vector machines-kernels and the kernel trick. *Notes*, 26(3):1–16, 2006.

[164] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

DOI:10.6814/NCCU202100261

[165] Souhaib Ben Taieb and Rob J Hyndman. A gradient boosting approach to the kaggle load forecasting competition. *International journal of forecasting*, 30(2):382–394, 2014.

[166] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[167] http://scikit-learn.org/stable/.

[168] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[169] https://social-media-prediction.github.io/MM17PredictionChallenge/index.html.

[170] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[171] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14, 2007.

[172] Jinna Lv, Wu Liu, Meng Zhang, He Gong, Bin Wu, and Huadong Ma. Multi-feature fusion for predicting social media popularity. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1883–1888, 2017.

[173] Xiaowen Huang, Yuqi Gao, Quan Fang, Jitao Sang, and Changsheng Xu. Towards SMP challenge: stacking of diverse models for social image popularity prediction. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1895–1900, 2017.

[174] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[175] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016.

DOI:10.6814/NCCU202100261

[176] Baoyuan Wu, Weidong Chen, Peng Sun, Wei Liu, Bernard Ghanem, and Siwei Lyu. Tagging like humans: Diverse and distinct image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7967–7975, 2018.

DOI:10.6814/NCCU202100261

Appendix A

Publications

Journals

1. **Fatma S. Abousaleh**, Wen-Huang Cheng, Neng-Hao Yu, and Y. Tsao, "Multimodal Deep Learning Framework for Image Popularity Prediction on Social Media," *IEEE Transactions on Cognitive and Developmental Systems*. ***(Accepted, early version available)***.

2. **Fatma S. Abousaleh**, Tekoing Lim, Wen-Huang Cheng, Neng-Hao Yu, M. Anwar Hossain, and Mohammed F. Alhamid, "A Novel Comparative Deep Learning Framework for Facial Age Estimation," *EURASIP Journal on Image and Video Processing*, no.1 (2016), p.47.

DOI:10.6814/NCCU202100261