Biometric refinement of datasets for facial age estimation

Bešenić, Krešimir

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Krešimir Bešenić

BIOMETRIC REFINEMENT OF DATASETS FOR FACIAL AGE ESTIMATION

DOCTORAL THESIS

Zagreb, 2024

University of Zagreb

FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Krešimir Bešenić

BIOMETRIC REFINEMENT OF DATASETS FOR FACIAL AGE ESTIMATION

DOCTORAL THESIS

Supervisors: Professor Igor Sunday Pandžić, Ph.D.
Assistant Professor Jörgen Ahlberg, Ph.D.

Zagreb, 2024

Sveučilište u Zagrebu

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Krešimir Bešenić

BIOMETRIJSKO UNAPREĐENJE SKUPOVA PODATAKA ZA PROCJENU DOBI IZ SLIKA LICA

DOKTORSKI RAD

Mentori: Prof dr. sc. Igor Sunday Pandžić
Izv. prof. dr. sc. Jörgen Ahlberg

Zagreb, 2024.

The doctoral thesis was completed at the University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Telecommunications and at the company Visage Technologies.

Supervisor: Professor Igor Sunday Pandžić, Ph.D.
Department of Telecommunications,
Faculty of Electrical Engineering and Computing,
University of Zagreb

Supervisor: Assistant Professor Jörgen Ahlberg, Ph.D.
Department of Electrical Engineering,
Computer Vision Laboratory,
Linköping University

The thesis has 135 pages.

Thesis number: _____

About the Supervisors

Igor Sunday Pandžić is a Professor at the Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. He leads the Human-Oriented Technologies Laboratory (HOTLab). He teaches undergraduate and postgraduate courses in the fields of virtual environments and communications. His main research interests are in the field of computer graphics and, more recently, computer vision, with a particular interest in face analysis and animation and a strong focus on applications of these technologies. Igor also worked on networked collaborative virtual environments, computer-generated film production, and parallel computing. He published five books and around 100 papers on these topics.

Jörgen Ahlberg received his M.Sc degree in Computer Science and Engineering in 1996 and his Ph.D in Electrical Engineering in 2002, both from Linköping University, Sweden. He then held positions as a scientist and research leader at FOI, the Swedish Defence Research Agency, for nine years. He is currently an Associate Professor at Linköping University as well as Oslo University. His research interests in the general area of image analysis and vision include analysis of facial images as well as automatic detection, classification, and tracking in thermal, hyperspectral, and radar systems. He has published more than 70 scientific papers, of which four are award-winning, and has been granted six patents.

O mentorima

Igor Sunday Pandžić redoviti je profesor na Zavodu za telekomunikacije Fakulteta elektrotehnike i računarstva na Sveučilištu u Zagrebu. Voditelj je istraživačkog laboratorija Human-Oriented Technologies Laboratory (HOTLab). Predaje preddiplomske i diplomske predmete u području virtualnih okruženja i komunikacije. Glavno područje njegovog istraživačkog interesa je računalna grafika uz dodatak računalnog vida sa posebnim naglaskom na analizu i animaciju lica te primjenu tih tehnologija. Igor je također radio na umreženim virtualnim okruženjima, računalno generiranoj filmskoj produkciji i paralelnom računarstvu. Objavio je pet knjiga i oko 100 članaka u spomenutim područjima.

Jörgen Ahlberg je diplomirao 1996. godine, a doktorsku disertaciju obranio 2002. godine na Sveučilištu u Linköpingu u Švedskoj. Nakon toga radio je devet godina kao znanstvenik i voditelj istraživanja u Švedskoj agenciji za obranu (FOI). Trenutno je izvanredni profesor na Sveučilištu u Linköpingu, kao i Sveučilištu u Oslu. Njegovi istraživački interesi iz područja analize slika i računalnog vida uključuju analizu slika ljudskog lica te automatsku detekciju, klasifikaciju i praćenje u termalnim, hiperspektralnim i radarskim sustavima. Objavio je više od 70 znanstvenih radova, od kojih su četiri dobila nagrade, i šest patenata.

Acknowledgments

I would like to express my gratitude to my supervisors, Dr. Igor S. Pandžić and Dr. Jörgen Ahlberg, for providing support throughout the entire process of writing this thesis, as their expertise and guidance were essential for this accomplishment. Moreover, I am thankful for their guidance in challenging life circumstances, helping me keep moving in the right direction. I would also like to thank Dr. Ivan Gogić for sharing some of his unwavering optimism and encouraging me to start this journey. I want to acknowledge Visage Technologies for allowing me to focus on my research without being burdened by financial constraints and express my appreciation to my FTD colleagues for countless insightful academic discussions. Finally, I owe a debt of gratitude to my family and friends, especially my parents, sisters, and girlfriend, for their love, emotional support, and for giving me all the time and space I needed to climb this mountain.

Abstract

This thesis explores the intricate task of facial age estimation with a focus on the role of data in the design of machine-learning-based systems. Supervised machine learning algorithms rely on large amounts of labeled data to learn a mapping between the input data (e.g., facial images or videos) and associated labels (e.g., age or gender). Due to difficulties related to biometric data collection, facial image datasets with biometric trait labels are scarce and frequently limited in terms of size and sample diversity. Web-scraping approaches for automatic data collection can produce large amounts of heterogeneous but weakly labeled and noisy data. The portion of erroneous samples in such data can be automatically reduced by the proposed unsupervised biometric data filtering method. The problem of data scarcity is further addressed by designing a new public facial image dataset. The proposed derived dataset is thoroughly compared to various state-of-the-art datasets with very favorable results and interesting insights about the significance of data. Video sequences can capture many spatiotemporal dynamic cues related to age progression. However, video-based age estimation is an understudied research field. This may be caused by a lack of well-defined and publicly accessible video benchmark protocols, as well as the absence of video-oriented training data. The former issue was addressed by proposing a carefully designed video age estimation benchmark protocol and making it publicly available. The latter issue was addressed by designing a video-specific age estimation method that leverages pseudo-labeling and semi-supervised learning. The experimental results indicate that using video data unlocks great potential in the age estimation research field.

Keywords: face analysis, age estimation, deep learning, data filtering, dataset design, semi-supervised learning, temporal models

Prošireni sažetak

Biometrijsko unapređenje skupova podataka za procjenu dobi iz slika lica

Doktorski rad fokusiran je na algoritme za procjenu ljudske dobi na temelju slika lica. Procjena dobi ima razna potencijalna područja primjene. Istraživanja pokazuju da ljudi prilikom procjene dobi u prosjeku griješe za 5 do 7 godina. Kako bi se tehnologija procjene dobi mogla uspješno primjenjivati u uvjetima stvarnog svijeta (*engl. in-the-wild*), modeli za automatsku procjenu dobi trebaju nadmašiti ljudske sposobnosti. Suvremeni algoritmi strojnog učenja omogućili su nadmašivanje ljudskih sposobnosti u srodnim područjima istraživanja kao što su raspoznavanje lica i klasifikacija objekata.

Brojna istraživanja ukazuju na to da nedostajuća karika za daljnje poboljšanje modela za procjenu dobi leži u podacima za učenje. Javno dostupni podaci za učenje nisu adekvatni. Ručno skupljani podaci za učenje su pouzdani, no sadrže relativno mali broj uzoraka i ne pokrivaju raznolikost uvjeta stvarnog svijeta u dostatnoj mjeri. Skupovi za učenje prikupljani s internet izvora (*engl. web-scraped*) imaju veliki broj uzoraka i dobru raznolikost, no imaju nepouzdane oznake dobi. Nadalje, većinski udio metoda za procjenu dobi oslanja se na korištenje pojedinačnih slika lica, no istraživanja pokazuju da korištenje video podataka za analizu lica krije veliki potencijal.

Kako bi se osiguralo da modeli imaju dobre sposobnosti generalizacije i dostatnu točnost procjene u stvarnim uvjetima, potrebno ih je evaluirati na testnim skupovima podataka koji u dobroj mjeri preslikavaju stvarne uvjete. Nadalje, potrebno je koristiti javno dostupne i precizno definirane evaluacijske protokole kako bi se dobiveni rezultati mogli reproducirati, a izvedeni zaključci verificirati.

Doprinosi doktorskog rada fokusirani su na rješavanje prethodno navedenih problema s ciljem poboljšanja rada sustava za automatsku procjenu dobi. Preciznije, usmjereni su na dizajn metoda za pročišćavanje skupova podataka za učenje, dizajn protokola za evaluaciju, istraživanje potencijala video podataka i učenje modela na neoznačenim podacima. Doktorski rad podijeljen je na sedam poglavlja, sažetih u nastavku.

Prvo poglavlje - Uvod

Ljudski vizualni i kognitivni sustavi omogućavaju nam da obavljamo vrlo kompleksne zadatke, često i na nesvjesnoj razini. Čak i nakon vrlo kratkog pogleda na nečije lice, u mogućnosti smo s lakoćom procijeniti odgovarajuću dobnu skupinu, spol, etnicitet, emocionalno stanje i mnoge druge značajke. U današnjem svijetu, digitalne kamere postaju sveprisutna pojava, a mogućnosti elektroničkih uređaja rastu impresivnom brzinom, time stvarajući plodno tlo za razvoj sustava računalnog vida. Već desetljećima, istraživačke skupine iz akademskih i industrijskih krugova ulažu velike napore kako bi replicirali spomenute ljudske sposobnosti metodama

računalnog vida. Ovaj doktorski rad usmjeren je na automatsku procjenu dobi iz dva glavna razloga. Prvo, tehnologija automatske procjene dobi ima brojna potencijalna područja primjene. Drugo, automatska procjena dobi predstavlja značajan istraživački izazov. Procjena dobi jedan od najzahtjevnijih problema u području automatske analize lica.

Potencijalna područja primjene automatske procjene dobi uključuju biometrijske sigurnosne sustave, sustave za interakciju čovjeka i računala, sustave za kontrolu pristupa, sustave za analizu prividne dobi, sustave za elektroničko upravljanje odnosima s kupcima i brojne druge. Unatoč velikom potencijalu, primjena ove tehnologije u mnogim od spomenutih područja ograničena je nedovoljnom preciznošću modela za procjenu dobi. Čak i ljudi značajno griješe kod procjene dobi nepoznatih osoba. Prema istraživanjima, ljudi u prosjeku griješe za 5 do 7 godina. Razlog za to leži u kompleksnosti procesa starenja i brojnim intrinzičnim i ekstrinzičnim faktorima koji utječu na njega. Intrinzični faktori uključuju faktore vezane uz ljudski organizam, kao što su spol, rasa, struktura mišićnog tkiva i kostiju, zdravlje te brojne druge genetičke predispozicije. Ekstrinzični faktori uključuju razne životne uvjete i navike kao što su izloženost suncu i vjetru, korištenje duhanskih proizvoda, prehrana te izloženost stresu. Uz to, estetska kirurgija, korištenje šminke i raznih drugih proizvoda za prikrivanje dobi u velikoj mjeri utječu na prividnu dob. Navedeni faktori otežavaju razvoj sustava za preciznu procjenu dobi iz slika lica.

Današnji sustavi za automatsku procjenu dobi temelje se na metodama strojnog učenja. Takve metode zahtijevaju velike količine podataka za učenje, no podaci koji uključuju slike lica i osobne informacije poput dobi i spola su teško dostupni. Brojna istraživanja ukazuju na to da su podaci za učenje najbitnija komponenta današnjih sustava za automatsku procjenu dobi. Nadalje, istraživanja ukazuju i da video podaci sadrže više informacija od pojedinačnih slika. Unatoč tome, korištenje videa za automatsku procjenu dobi nije dovoljno istraženo, prvenstveno zbog manjka podataka za učenje i nedostupnosti precizno definiranih javnih protokola za evaluaciju. Glavni doprinosi ovog rada, usmjereni na rješavanje tih problema, su slijedeći:

- •Nenadzirana biometrijska metoda za filtriranje podataka iz slikovnih baza lica i njezina primjena na *state-of-the-art* baze za procjenu dobi.
- •Strategija za podržavanje i unapre đenje procesa povezivanja baza lica temeljena na biometrijskom filtriranju, koja rezultira u novoj izvedenoj bazi nazvanoj *Biometrically Filtered Famous Figure Dataset* (B3FD).
- •Analiza utjecaja korištenja predstavljene B3FD baze lica na pet metoda za procjenu dobi.
- •Protokol za procjenu dobi temeljen na video zapisima (*CCMiniVID*) koji omogućuje adekvatnu usporedbu metoda temeljenih na video zapisima u stvarnim uvjetima i kroz više skupova podataka.
- •Djelomi čno nadzirana metoda procjene dobi temeljena na video zapisima koja nadilazi nedostatak označenih podataka za učenje i nadmašuje referentnu metodu baziranu na slikama.

Drugo poglavlje - Osnove strojnog učenja

Strojno učenje jedan je od centralnih problema u području umjetne inteligencije. To je grana računalne znanosti koja proučava algoritme i statističke modele koji mogu izvršavati određene zadatke bez eksplicitnog programiranja. Drugim riječima, algoritmi strojnog učenja su algoritmi koji imaju mogućnost učenja iz podataka. Drugo poglavlje daje pregled osnovnih koncepata i algoritama iz područja strojnog učenja koji se direktno koriste u ovom radu.

Algoritmi strojnog učenja mogu se kategorizirati na različite načine, a granice između tih kategorija je ponekad teško precizno definirati. Glavne kategorije algoritama strojnog učenja su nadzirano učenje, nenadzirano učenje i podržano učenje. Djelomično nadzirano učenje kombinira komponente nadziranog i nenadziranog učenja.

Suvremeni sustavi za automatsku procjenu dobi često su temeljeni na nadziranom učenju neuronskih mreža. Neuronske mreže dizajnirane su za učenje direktno iz podataka, bez potrebe za ručnom izradom algoritama za izlučivanje značajki. Konvolucijske neuronske mreže su popularna kategorija neuronskih mreža posebno dizajniranih za rade sa slikovnim podacima. Duboke konvolucijske neuronske mreže osnova su dubokog učenja. Metode dubokog učenja koriste veliki broj jednostavnih nelinearnih modula za učenje značajki na više razina, pri čemu se svakim dodanim modulom uče značajke na sve višoj razini apstrakcije, što ujedno omogućuje i efikasno učenje vrlo kompleksnih nelinearnih funkcija.

Metode za procjenu, kako dobi tako i brojnih drugih atributa, mogu se podijeliti na klasifikacijske, regresijske i napredne metode. Cilj klasifikacijskih metoda je određivanje funkcije preslikavanja ulaznih podataka na konačni skup kategorija oznaka, regresijske metode preslikavaju ulazne podatke na oznake s kontinuiranim numeričkim vrijednostima, dok velik broj metoda kombinira klasifikacijski i regresijski pristup na napredne načine.

Uz algoritme strojnog učenja, arhitekture neuronskih mreža i metode za procjenu, veliku ulogu u području strojnog učenja imaju sami podaci. Podaci za procjenu dobi sastoje se od slika lica i pripadnih oznaka dobi. Drugo poglavlje zaključeno je diskusijom o značaju podataka, uključujući podatke za predučenje, učenje i testiranje. Neki od glavnih problema u području procjene dobi proizlaze iz prikupljanja slika u suviše kontroliranom okruženju, netočnih oznaka dobi i nebalansiranih demografskih distribucija.

Treće poglavlje - Automatska analiza lica

Kod velikog dijela metoda za automatsku analizu lica mogu se prepoznati četiri glavna slijedna koraka. To su detekcija lica, predprocesiranje, izlučivanje značajki i predikcija. Detekcija lica rezultira informacijom o poziciji i veličini lica u slici. Cilj predprocesiranja je pripremiti podatke u konzistentnom i što informativnijem obliku za nadolazeće korake. Najčešće metode predprocesiranja uključuju izrezivanje lica iz slike, promjenu veličine i normalizaciju vrijed-

nosti u slici. Izlučivanje značajki je korak u kojem je cilj izlučiti informativne, diskriminativne i kompaktne vektore značajki koji se mogu učinkovito koristiti u posljednjem koraku, odnosno predikciji. Cilj samog koraka predikcije je odrediti ciljani atribut ulaznog uzorka (npr. dob ili spol) na temelju izlučenog vektora značajki. Predikcija se najčešće provodi metodama klasifikacije ili regresije. Važno je napomenuti da se koraci izlučivanja značajki i predikcije u suvremenim metodama baziranim na neuronskim mrežama i dubokom učenju najčešće optimiziraju zajednički prilikom procesa učenja.

Zadatak automatske procjene dobi može se formulirati na tri glavna načina. To su procjena dobne skupine, procjena kronološke dobi i procjena prividne dobi. Procjena dobne skupine najjednostavniji je zadatak jer je potrebno odrediti okvirnu dob i kategorizirati uzorak u konačni i relativno mali broj dobnih skupina. Osnovni primjer je binarna klasifikacija maloljetnih i odraslih osoba. U literaturi se najčešće provodi kategorizacija na 7 ili 8 dobnih skupina. Zadatak procjene kronološke dobi najzahtjevniji je jer zahtjeva određivanje precizne numeričke vrijednosti koja predstavlja stvarnu dob osobe. Procjena prividne dobi također zahtijeva određivanje precizne numeričke vrijednosti, no ta vrijednost predstavlja prividnu dob baziranu na izgledu osobe.

Četvrto poglavlje - Povezani radovi

Automatska procjena dobi iz slika lica istražuje se već tri desetljeća. Ovo plodno istraživačko područje rezultiralo je velikim brojem predloženih algoritama, metoda i skupova podataka. Kako bi se proveo sustavni pregled područja, predložena je taksonomija koja organizira dizajn sustava za procjenu dobi prema vrsti ulaznih podataka, tipu značajki i algoritmu procjene. Iako većina metoda koristi pojedinačne slike kao ulazne podatke, određeni broj metoda koristi video sekvence umjesto slika. Veliki broj metoda baziran je na ručno dizajniranim značajkama, dok se većina suvremenih metoda bazira na učenim značajkama, često dobivenim konvolucijskim neuronskim mrežama. Najčešće korišteni algoritmi procjene su regresija i klasifikacija, no u literaturi se navodi i veliki broj naprednijih algoritama za procjenu kao što su redno rangiranje i učenje distribucija. Uz velik broj metoda za procjenu dobi, literatura navodi i brojne skupove podataka. Slikovni skupovi podataka su značajno zastupljeniji u odnosu na video skupove podataka, kojih je javno dostupno samo nekoliko. Glavni zaključci pregleda područja navedeni su u nastavku.

Konsolidirano je uvjerenje da duboko učenje nudi ogromne prednosti u odnosu na metode temeljene na ručno izrađenim značajkama i klasičnim algoritmima učenja. Istraživanja pokazuju da korištenje video podataka za procjenu dobi krije veliki potencijal, no ta grana istraživanja nije dovoljno istražena. Potencijalni uzrok ovog problema je nedostatak javno dostupnih podataka za učenje, kao i precizno definiranih protokola za evaluaciju. Iako se značajan udio literature bavi razvojem samih metoda za procjenu, konsenzus oko najprikladnije metode za procjenu

dobi nije postignut.

Novija istraživanja ističu brojne probleme vezane uz postojeće javno dostupne evaluacijske protokole. Glavni problemi proizlaze iz pristranosti skupova podataka za evaluaciju i neprecizno definiranih protokola. Uz to, doprinos svih bitnih komponenti u sustavu najčešće nije definiran. Validnost konvencionalnih usporedbi metoda za procjenu dobi stoga se sve češće dovodi u pitanje. Nova istraživanja ukazuju na to da prikladnost i veličina skupa podataka za učenje igraju najbitniju ulogu u dizajnu sustava za procjenu dobi. Pokazano je i da arhitektura neuronske mreže i način predprocesiranja slika lica mogu značajnije doprinositi od odabira same metode za procjenu dobi. Iz tih razloga, doprinosi ovog rada fokusirani su prvenstveno na probleme iz područja procjene dobi vezane uz same podatke.

Peto poglavlje - Nenadzirano biometrijsko filtriranje podataka za unapređenje procjene dobi

U usporedbi s ručno prikupljenim skupovima podataka za procjenu biometrijskih atributa iz slika lica, skupovi podataka dobiveni automatskim metodama prikupljanja s interneta (*engl. web-scraped*) daleko su napredniji u pogledu količine i raznolikosti uzoraka, no karakterizira ih i nepouzdanost oznaka.

Predložena metoda za nenadzirano biometrijsko filtriranje podataka može automatski smanjiti broj pogrešnih uzoraka u skupovima podataka prikupljenim s interneta, kombiniranjem samo nekoliko općih algoritama. Temeljna ideja predložene metode je automatsko grupiranje slika sa sličnim biometrijskim značajkama u grupe slika iste osobe, te zadržavanje samo najveće grupe. Implementirani filtracijski sustav rezultirao je drastičnim smanjenjem broja uzoraka na dva najznačajnija skupa podataka prikupljenih s interneta, pri čemu je odbačeno do 52% uzoraka. Detaljno testiranje pokazalo je da modeli koji se temelje na filtriranim podacima u značajnoj mjeri nadmašuju modele koji se temelje na sirovim i konvencionalno obrađenim podacima, što ukazuje na manji broj pogrešno označenih uzoraka i poboljšanu dosljednost oznaka. Testiranje sposobnosti generalizacije na podacima prikupljenim u stvarnim uvjetima također je pokazalo da raznolikost podataka nije narušena, unatoč drastičnom smanjenju broja uzoraka.

Rezultati dobiveni predloženom metodom filtriranja prošireni su dodatnom strategijom biometrijskog filtriranja koja je osmišljena kako bi ojačala i usavršila proces spajanja javno dostupnih skupova podataka prikupljenih putem interneta. Predloženi proces spajanja triju različitih izvora podataka prikupljenih putem interneta rezultirao je novim slikovnim skupom podataka. Predloženi B3FD skup podataka (*engl. Biometrically Filtered Famous Figure Dataset*) eksperimentalno je validiran i uspoređen s trenutno najboljim javno dostupnim skupovima podataka. B3FD je dosljedno nadmašio sve evaluirane skupove podataka te je učinjen javno dostupnim.

Opsežna eksperimentalna evaluacija istaknula je važnost kvalitete podataka i dosljednosti oznaka. Rezultati modela treniranih na podskupovima podataka dobivenim predloženim meto-

dama filtriranja nadmašili su rezultate modela treniranih na većim skupovima podataka. Dodatno, pokazano je da korištenje prikladnijih podataka za učenje doprinosi više od korištenja naprednijih metoda za procjenu dobi.

Šesto poglavlje - Put prema procjeni dobi temeljenoj na video podacima

Šesto poglavlje usredotočeno je na rješavanje dva istaknuta problema u području automatske procjene dobi iz video podataka. To su nedostatak precizno definiranog i javno dostupnog evaluacijskog protokola za metode temeljene na video podacima te nemogućnost nadziranog učenja modela za procjenu dobi iz video podataka zbog nedostatka označenih podataka za učenje.

Kako bi se riješio prvi od spomenutih problema, dizajniran je novi evaluacijski protokol baziran na video podacima. Postojeći javno dostupni skup video podataka proširen je meta-podacima dobivenim komercijalnim sustavom za praćenje lica i pročišćen polu-automatskim filtriranjem. Predloženi protokol učinjen je javno dostupnim, zajedno sa svim potrebnim meta-podacima i odgovarajućom programskom podrškom.

U svrhu rješavanja drugog spomenutog problema u ovom području, predložena je nova metoda koja svladava problem nedostupnosti označenih podataka za učenje te ujedno poboljšava točnost modela za procjenu dobi. Predložena metoda za procjenu dobi iz video podataka temeljena je na pseudo-oznakama i djelomično nadziranom učenju. Predloženi model kombinira 2D konvolucijsku neuronsku mrežu za izlučivanje značajki i temporalnu konvolucijsku mrežu za temporalno modeliranje. Testiranja na predloženom protokolu za evaluaciju na video podacima pokazala su da je greška procjene dobi umanjena za više od 15%, dok je stabilnost procjene poboljšana za čak više od 50%. Time je pokazano da je model dobiven djelomično nadziranim učenjem na video podacima nadmašio referentni model učen na slikama, odnosno model koji je korišten za dobivanje samih pseudo-oznaka.

Sedmo poglavlje - Zaključci

Procjena kronološke dobi iz slika lica jedan od najkompleksnijih problema u području automatske analize lica. Glavni razlog za to je personalizirana i stohastička priroda procesa starenja na kojeg utječu mnogi intrinzični i ekstrinzični faktori. Kako bi se sustavi za automatsku procjenu dobi mogli uspješno koristiti u brojnim potencijalnim područjima primjene, potrebno ih je usavršiti do razine naprednije od ljudskih sposobnosti.

Glavni doprinosi ovog rada usmjereni su na podatke za učenje i evaluaciju modela za procjenu dobi. Točnost i robusnost modela strojnog učenja uvelike ovise o količini i kvaliteti korištenih podataka za učenje. Valjanost i provjerljivost rezultata istraživanja izravno ovise o dostupnosti i kvaliteti javnih protokola za evaluaciju. Brojna istraživanja ukazuju na postojanost problema u javno dostupnim skupovima podataka. U eri naprednih algoritama strojnog

učenja, kompleksnih modela i impresivnih mogućnosti elektroničkih uređaja, ključ za razvijanje sustava za procjenu dobi koji su zaista robusni u uvjetima stvarnog svijeta leži u samim podacima.

Ključni pojmovi: analiza lica, procjena dobi, duboko učenje, filtriranje podataka, dizajn skupova podataka, djelomično nadzirano učenje, temporalni modeli

Contents

Chapter 1

Introduction

Human visual and cognitive systems allow us to perform many incredibly complex tasks effortlessly. By taking even the briefest look at someone's face, we can extract a plethora of information. Who is this person? Does this person look like a male or a female? How old does this person look like? What is the ethnicity of this person? How does this person feel? We can often answer most of these questions with ease [1]. This type of information plays a pivotal role in face-to-face social interactions [2].

Today's omnipresence of digital cameras, along with recent advancements in the computational capabilities of electronic devices, provide fertile ground for the development of advanced computer vision systems. For more than three decades, researchers have aspired to replicate human face analysis capabilities with automated, vision-based systems. Not only does this make for an interesting research field that helps us to understand human nature better, but accurate and robust automatic age estimation can enable a variety of application fields, such as human-computer interaction, access control, electronic customer relationship management, and biometric security [3].

Aging is a personalized, stochastic, uncontrollable, inevitable, and irreversible process [4]. It causes visually observable changes to the human body that enable us to estimate a subject's age based solely on visual cues. Facial chronological age estimation is the process of estimating the number of years elapsed from one's birth to a certain point in life using visual artifacts on the face [4]. This is one of the most challenging tasks from the face analysis research field [1] due to the personalized nature of the aging process and many intrinsic and extrinsic factors that influence it [5]. Despite years of research and devotion from both academia and industry towards aging process modeling, estimation algorithm design, data collection, and defining testing protocols, we are yet to reach age estimation accuracy sufficient for many of the advanced application fields [2, 4].

1.1 Motivation

The focus of this thesis is human age estimation based on visual facial information. Cues useful for age estimation can be extracted from many different physical characteristics and behavioral traits, such as teeth [6], fingerprints [7], bones [8], hands [9], irises [10], voice [11], gait [12], and even head movement [13] and keystroke dynamics [6]. While humans often combine modalities to estimate the age of other people, this thesis focuses on the single modality that provides the most visual cues related to aging progression and can enable unintrusive, automated, computer-vision-based estimation even from a single image or video frame.

Visual analysis of human faces can facilitate the estimation of many other useful information, such as identity [14], facial expression [15], eye gaze [16], gender [17], ethnicity [18], and numerous other soft biometric traits [19]. Our motivation to focus on age estimation is twofold. As a famous person once said, we are not doing this because it is easy but because it is hard. Age estimation is one of the most challenging computer vision tasks [20]. Reliable, unintrusive, and automated age estimation is the basis for many interesting application fields with vast potential. We explore these statements in the following sections.

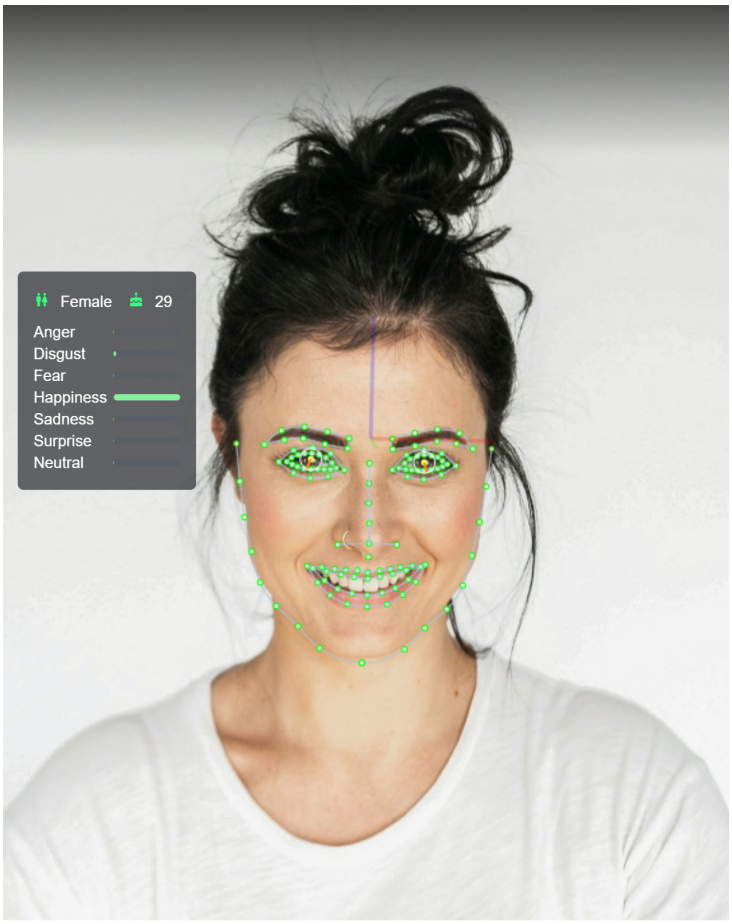

Figure 1.1: Commercial application of face analysis algorithms by the company Visage Technologies.

1.1.1 Application fields

Contemporary computer vision systems enable automatic face analysis from images or videos, often with real-time processing capabilities. Automatic age estimation technology has many application scenarios, either as a standalone feature or as a supporting feature in more complex face analysis systems. Figure 1.1 shows an example of a commercial application of face analysis technology.

Biometrics and security

Age is a soft biometric attribute. Biometrics refers to the problem of subject identification based on a certain unique physical characteristic. Biometric systems can be based on face, iris, hand geometry, and fingerprint recognition [21]. Soft biometric attributes are traits such as gender, height, and eye color that provide some useful information about the subject but are not distinctive enough for identification [17]. The intrusiveness of biometric systems based on fingerprint or iris recognition reduces their applicability compared to systems based on facial image analysis that do not require physical contact, subject cooperation, or attention. The accuracy and robustness of biometric systems, often used in security applications, can be improved by the use of age estimation [4, 21]. The security of many everyday systems can also be improved by the direct use of automatic age estimation: age can be useful information in ATM fraud monitoring [21], vehicle in-cabin systems can be designed to monitor if a child is left unattended by mistake [22], and surveillance camera footage processed by an age estimation system can be used in demographics analysis for crime prevention [22].

Human-Computer Interaction (HCI)

Around 93% of information in human communication is conveyed nonverbally [23]. Face-based Human-Computer Interaction systems can be designed to leverage this nonverbal information. HCI is rapidly becoming an omnipresent phenomenon in our daily lives. HCI interface, vocabulary, and services can automatically be selected to best suit the users' age because many preferences, such as linguistics, aesthetics, and consumption habits, change with age [23].

Access control

Many products and services require access control based on customer age. Automatic age estimation can be useful for underage purchasing restrictions of alcoholic and tobacco products [23, 24]. Access to certain age-restricted websites can automatically be managed [23]. TV sets equipped with cameras, such as contemporary smart TVs, can limit streaming outputs to prevent children from accessing unwanted content [4]. Access to dangerous or inappropriate rides can be automatically regulated in amusement parks [25].

Figure 1.2: Access control system (right) in a smart store (left) preventing minors from entering the section containing age-restricted products (i.e., alcoholic beverages).

Electronic customer relationship management (ECRM)

Businesses often benefit from a personalized approach to customers. Preferences and expectations related to products and services change with age [26]. Market trends can be monitored so that products and services are customized to accommodate different age groups [4]. Electronic customer relationship management (ECRM) leverages algorithmic approaches for establishing individualized relationships with customers [25]. Age estimation can play an important role in such a management strategy: instead of building complex customer databases and storing sensitive personal data, customer relationships can be improved by estimating age information in real time, during the interaction with the customer.

Other use cases

The employment process can be enhanced by automatically estimating the age of recruits [4]. Age estimation can help to detect victims and suspects during investigations of child sex abuse materials [27]. Identification of missing persons can be aided by age simulation systems [4]. Applications for apparent age estimation, as reviewed in [28], include medical diagnosis [29, 30], facial beauty product development [31, 32], movie role casting [31, 32], and analysis of the effects of plastic surgery [21, 33] and anti-aging treatments [29, 32].

1.1.2 Challenges

As the applicability of age estimation systems is evident from the review in the previous section, we continue by analyzing challenges that make the design of accurate age estimation systems such a difficult task. Human visual and cognitive systems allow us to perform many incredibly complex tasks effortlessly. However, estimating chronological age from unknown faces based only on visual cues is a challenging task, even for humans. Han et al. [34] carried out crowdsourcing experiments with 10 workers on FG-NET [35] and PSCO [34] age estimation benchmark datasets. On average, they misestimated age by 4.7 and 7.2 years, respectively. Gang et al. [36] performed a similar experiment with 29 volunteers on a subset of the FG-NET benchmark dataset, resulting in an error of 6.23 years on average. Agustsson et al. [30] experimentally tested *the wisdom of the crowd* by using crowdsourcing and collecting nearly 300,000 votes on 7,591 images from their Appa-Real benchmark. 38 votes per image were cast on average, resulting in an apparent age label for each of their images. Those labels were wrong by 4.57 years on average on their test set, compared to the actual chronological age labels. These error rates are too high for many of the aforementioned application fields, meaning that automatic age estimation systems should aim to surpass human performance by a considerable margin. The following sections cast some light on why age estimation is so challenging, both for humans and machines.

Figure 1.3: The influence of sun exposure on aging progression. The images show twins that are 61 years old. The twin on the right side had $\approx$ 10 hours per week more sun exposure. The difference in the perceived age is 11.25 years. Reprinted from [37] with permission from Wolters Kluwer Health, Inc.

Facial aging progression

The human face conveys many features that can indicate aging progression, such as skin texture (wrinkles, age spots, freckles), bone structure (face shape, cranium size, yaw shape), and facial hair (presence, amount, and color) [6]. The nature of the aging process, which causes changes in those features, is stochastic, uncontrollable, inevitable, and irreversible [4]. We age from the moment we are born to the end of our lives. In childhood, the most obvious facial appearance changes are related to craniofacial growth, as the cranium shape changes from circular to oval. The forehead slopes back, creating more space for the mouth, nose, eyes, and ears to expand, while the chin becomes more protrusive [4]. In adulthood, some minor face shape changes are observable, but the most obvious aging progression indications come from skin texture changes [4]. The natural aging process involves a decrease in collagen and elastin production, the disappearance of fat cells, skin losing the ability to retain moisture, lower turnover of new skin cells, and slowdown of dead skin cell shedding [38]. The deterioration of facial bones also affects the skin texture [39]. All this makes the skin less elastic, thinner, and sagging, causing wrinkles and skin to appear leathery. Damage to melanin-producing cells, happening over time primarily due to UV rays, causes freckles, age spots, and non-uniform skin tone [4, 39].

Figure 1.4: The influence of smoking on aging progression. The images show twins that are 52 years old. The twin on the left side had 20 years longer smoking history. The difference in the perceived age is 6.25 years. Reprinted from [37] with permission from Wolters Kluwer Health, Inc.

Some of the major influences on aging progression are gravity, exposure to ultraviolet (UV) rays, facial muscular activities, bone restructuring, and maturing of soft tissue [4, 40]. These factors are usually categorized into intrinsic or extrinsic [4, 5, 22, 41]. We review them in more detail in the following sections, while the effects of some of these factors are depicted in Figures 1.3, 1.4, 1.5, 1.6, and 1.7.

Intrinsic aging factors

Intrinsic aging factors are related to the human body and its nature, such as genetics, bone structure, muscular activity, gender, and race. Genetic makeup is the primary dependency of individual aging patterns [42]. Gender, race, and facial expressions directly affect aging patterns [43]. Male aging pattern differences include changes in facial hair, facial vascularity, sebaceous content, and bone and fat absorption rates, while females tend to develop more deep wrinkles around the eyes as their skin contains a far lower number of appendages [38]. Muscular activity related to some facial expressions, such as smiling, frowning, and surprise, can cause wrinkle-like changes on facial skin [4]. Certain skin diseases can change facial skin appearance and make age estimation more difficult [41].

Extrinsic aging factors

Extrinsic aging factors include factors external to the human body that influence aging progression, such as environment, lifestyle, and occupation. Environmental factors that advance aging progression include exposure to UV rays [44], wind [45], and air pollution [46]. Some of the lifestyle and occupational factors that can expedite aging are smoking, drug use, and psychological stress [45]. Even an unhealthy diet can haste aging [4], while scratches and burns may appear as signs of aging [41]. This means that aging can manifest itself very differently based on the subject's life choices, circumstances, and environment, even for individuals with very similar genetic and demographic predispositions.

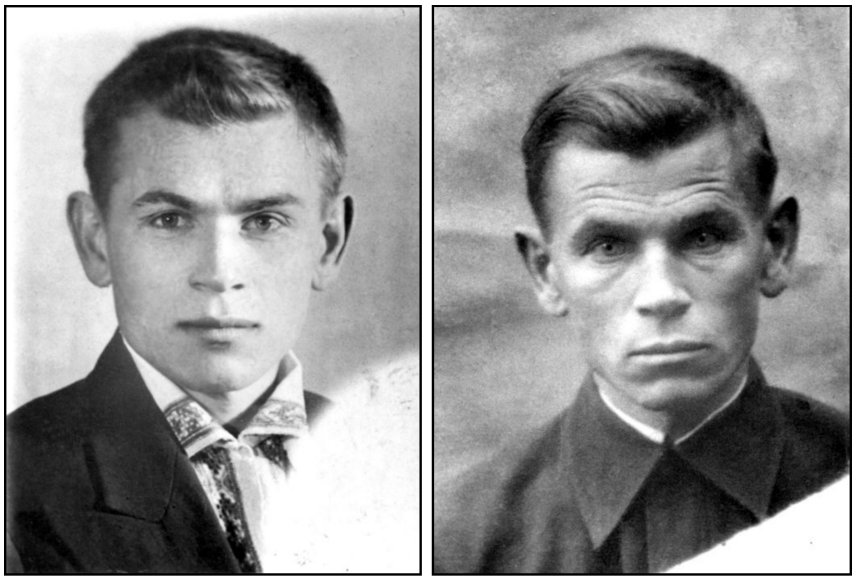

Figure 1.5: The influence of psychological stress and malnutrition on aging progression. The images show a soldier before and after 4 years of war (1941-1945). The images are exhibited in the Andrei Pozdeev Museum (http://mus-pozdeyev.ru/p82).

Appearing younger

Aging is an irreversible process. However, people often use anti-aging products and treatments to appear younger than they really are. Makeup, beauty care products, and cosmetic surgeries can decrease signs of facial aging [25]. This makes visual age estimation even more challenging [22]. Makeup is often used to alter a person's perceived age by concealing age spots, brightening wrinkle-induced shadows, and highlighting and coloring cheeks [47]. Aesthetic facial surgery causes a significant and consistent reduction in perceived age, as a study shows that perceived age in plastic surgery patients was reduced by more than 7 years on average [48].

Figure 1.6: The influence of facial makeup on perceived age. Reprinted from [25] with permission from the author.

Imaging and data challenges

Not all automatic age estimation challenges are directly related to facial appearance. For an age estimation system to work correctly, input images must be of sufficient quality. The texture, wrinkles, and other visual cues can be lost due to diminished image quality [50]. Some of the usual causes of insufficient image quality are camera distance, low camera resolution, motion blur, inadequate lighting, and occlusions. All these are present in most real-world use case scenarios.

Even when the system receives high-quality inputs, it can easily fail if it is not trained using a sufficient number of samples covering all age groups and all the aforementioned aging variations. Due to challenges related to the collection of data with precise age labels, it is very difficult to cover all variations with appropriate and sufficient samples, even when using the internet [3, 22]. This causes most age estimation benchmark datasets to have imbalanced distributions [3].

Figure 1.7: The influence of facial rejuvenation surgery on perceived age. The images show a 42-year-old patient before and after the surgery. The difference in the perceived age is 14.2 years. Reprinted from [49] with permission from Elsevier.

1.2 About the thesis

This thesis is an outcome of years of academic research made in collaboration with the industry. This work was supported by the face tracking and analysis company Visage Technologies. While academic curiosity motivated us to review many novel, advanced, and experimental methods, industry insights guided us to focus on aspects of system development that are often overlooked in purely academic work: model generalization capabilities and performance in real-world conditions.

While our initial efforts were directed towards the development of advanced age estimation methods, with a special focus on the design of novel neural network architectures and age estimation algorithms, we repeatedly encountered two big obstacles. The first one was the lack of sufficient and adequate training data. In the literature review and the subsequent experimental sections, we argue that training data is the most important component in the design of contemporary age estimation systems. The second frequently emerging obstacle was the lack of a properly designed cross-dataset benchmark protocol, required to fairly compare different systems and their abilities to generalize in real-world conditions. In addition to the data-related obstacles, we observed that the majority of age estimation research is focused on image-based approaches. Unlike still images, video sequences offer spatiotemporal dynamic information that contains many cues related to age progression. The contributions of this thesis are focused on solving the two aforementioned data-related obstacles while also exploring the potential of videos and video-based methods for the refinement of automatic facial age estimation.

1.2.1 Contributions

Many hurdles in the facial age estimation field are directly related to training data issues. Web-scraping methods can generate very large but noisy datasets. Large amounts of erroneously labeled samples present in the frequently used web-scraped facial image datasets can impair performance of face analysis algorithms. The first objective of our research was to refine the performance of face analysis methods using a data-driven approach based on the automatic filtering of facial image datasets. An unsupervised biometric filtering method was developed and used to create improved versions of relevant public web-scraped datasets. The main idea of that method is to automatically group images with similar biometric metadata into clusters of images of the same person, and only to keep the largest cluster. The second objective of our work was to create a new age estimation image dataset suitable to train deep learning models. We proposed a biometric filtering strategy to reinforce and refine the merging process of multiple facial datasets and derived the new Biometrically Filtered Famous Figure Dataset (B3FD). We demonstrated B3FD's superiority over existing state-of-the-art age estimation datasets with respect to both real and apparent age estimation performance in the cross-dataset setting. We made the dataset publicly available. The final objective of our research was to explore the options for leveraging spatiotemporal dynamic cues for age progression extracted from video streams. We designed a new video-based age estimation benchmark from an existing public dataset, extended the metadata using a commercial face tracking system, and made the benchmark protocol, metadata, and video processing framework publicly available. To overcome the lack of labeled training data, we designed a semi-supervised video age estimation method that relies on pseudo-labeling. The proposed method outperformed its image-based counterpart, thus setting baseline results on the proposed benchmark. The main contributions of this work are summarized as follows:

- An unsupervised biometric data filtering method for facial image datasets and its application to state-of-the-art age estimation datasets.
- A biometric filtering strategy for the reinforcement and refinement of the merging process of facial datasets resulting in the newly derived Biometrically Filtered Famous Figure Dataset (B3FD).
- Analysis of the impact of using the proposed B3FD dataset on five age estimation methods.
- A video-based age estimation benchmark protocol (CCMiniVID) that enables fair comparison of video-based methods in real-world conditions and cross-dataset settings.
- A semi-supervised video-based age estimation method that overcomes the lack of labeled training data and outperforms the image-based baseline.

1.2.2 Organization of the thesis

The rest of the thesis is organized as follows. Chapter 2 introduces fundamental machine learning concepts used in the thesis, such as learning methods, estimation methods, and the concept of deep learning, while also reflecting on the significance of data in machine learning. Chapter 3 introduces the basics of the general face analysis framework, with a special focus on the age estimation framework and relevant metrics. Related work, including image-based and video-based datasets and methods, is reviewed in Chapter 4. The proposed unsupervised biometric data filtering method and the newly derived image-based age estimation dataset are described in Chapter 5, while Chapter 6 presents the new video-based benchmark protocol and the proposed semi-supervised video-based age estimation method. Finally, Chapter 7 concludes the findings of this thesis.

Chapter 2

Machine learning foundations

Machine learning is a subfield and one of the central problems of Artificial Intelligence (AI). It is a branch of computer science that studies algorithms and statistical models that can perform a specific task without explicit programming [51]. In other words, machine learning algorithms are algorithms that can learn from data [52]. In some ways, the goal of machine learning algorithms is to mimic the human learning process. From the algorithmic perspective, the learning process needs to be formally defined. Referring to [53], let us assume experience E, a class of tasks T, and performance measure P. A model is said to learn from experience E if its performance at task T, measured by P, improves with experience E. As stated in [52], it is difficult to formally define the wide variety of possible experiences, tasks, and performance measures. Giving an exhaustive overview would surpass the scope of this thesis. Instead, this chapter provides a brief overview of fundamental machine learning concepts and specific algorithms used throughout the thesis, while extensive overviews of the field are given in [52, 53, 54, 55].

2.1 Learning algorithms

While machine learning algorithms can be categorized in many different ways and while it is often difficult to define clear boundaries between the categories, a consolidated approach is to classify learning algorithms into Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Along with the main three categories, there are many popular hybrid learning algorithms, such as Semi-Supervised Learning, Self-Supervised Learning, and Multi-Instance Learning. However, the boundaries of these categories can be somewhat unclear. For example, according to [52], Supervised Learning and Unsupervised Learning are not formally defined terms and some machine learning technologies can be utilized to perform both of these tasks. Nevertheless, we continue by introducing some common formal definitions for the three basic learning algorithms used in our work.

2.1.1 Supervised learning

Supervised learning algorithms aim to learn a function that maps inputs to outputs based on a set of predefined exemplary input-output pairs [51]. In this setting, the outputs are often referred to as targets or labels. The term *supervised* originates from the view that the labels are provided by a teacher or instructor who is therefore supervising the learning process [52]. Referring to [56], we define supervised learning as follows. Assume availability of a dataset $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i \in 1,\ldots,N}$, consisting of N pairs of d-dimensional inputs $\mathbf{x}_i \in \mathbb{R}^d$ and labels $y_i \in Y$, where Y can represent a finite set of categories ($Y \in \{0, 1, \ldots, K\}$) or continuous values ($Y \in \mathbb{R}$). Based on the dataset $\mathscr{D}$, the supervised algorithm aims to find a function $f : \mathbb{R}^d \to Y$ that performs the mapping of inputs to outputs and can be used to correctly predict outputs from previously unseen inputs. This function can be defined as $y_i = f(\mathbf{x}_i)$, and the ability to perform beyond the training set $\mathscr{D}$ is called generalization. Generalization capabilities are evaluated on a test set: samples that do not appear in the training set $\mathscr{D}$, but come from the same underlying distribution. It is common practice to use only a part of the initial set $\mathscr{D}$ for training, while the rest of the samples are held out for validation and testing.

The two most common supervised learning algorithms are classification and regression. The choice depends on the learning task T and the type of labels $y_i \in Y$ in the dataset $\mathscr{D}$. Classification is used to map inputs to a finite set of label categories ($Y \in \{0, 1, \ldots, K\}$), while regression is used to map inputs to continuous values ($Y \in \mathbb{R}$).

2.1.2 Unsupervised learning

While supervised learning algorithms require a dataset $\mathscr{D}$ defined in a way that every sample x_i has a corresponding label y_i, unsupervised learning algorithms can extract information directly from samples x_i, without relying on labels. These algorithms are left to their own devices to discover and present the underlying data structure and interesting patterns [51]. Without labels, the learning process is not guided by a supervisor, hence making *unsupervised learning* a fitting name for this category of algorithms.

According to [54], the most common use cases of unsupervised learning are clustering, density estimation, and visualization. The goal of clustering is to discover groups of similar data, density estimation aims to determine the distribution of data within the input space, while the purpose of visualization algorithms is to project the data from a high-dimensional space down to two or three dimensions, aiding interpretability.

Of special focus for our work are clustering algorithms. The main principle of clustering is to automatically create groups of samples based on certain similarity criteria in such a way that the elements within the same group are similar while also being dissimilar to elements of other groups [56]. The K-means algorithm is one of the simplest and frequently used unsupervised

learning algorithms. Referring to [52, 56], we define it as follows. Given the number of clusters k and the training samples x_i, the algorithm initializes k centroids $\{\mu_1, \ldots, \mu_k\}$ with different random training samples. Each of the remaining samples is assigned to cluster j, where j is the index of the most similar centroid μ_j. Then, each centroid μ_j is updated to the mean of all training examples x_i assigned to the cluster j. These steps are repeated until the centroids stop changing.

Some other relevant examples of unsupervised learning algorithms are Principal Component Analysis (PCA) [57], Chinese Whispers Clustering [58], Affinity Propagation Clustering [59], and Mean Shift Clustering [60].

2.1.3 Semi-supervised learning

Semi-supervised learning can be seen as a combination of supervised and unsupervised learning [51]. In general, information associated with one of these tasks is used to improve the performance of the other [61]. For example, in a supervised classification setting, additional unlabeled data can be used to aid the classification process; in an unsupervised clustering setting, the clustering procedure might benefit from the information that some samples belong to the same class [61].

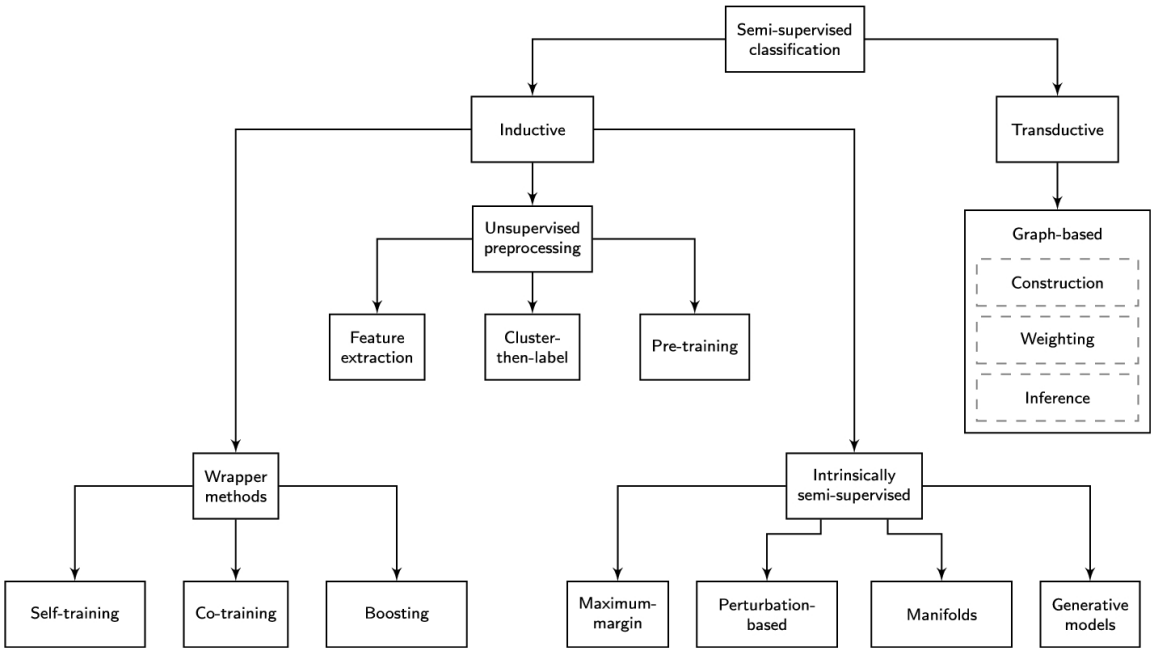

Figure 2.1: Taxonomy of semi-supervised classification. Reprinted from [61] under CC BY-SA 4.0 license.

A very practical aspect of semi-supervised learning is the ability to utilize unlabeled data in the supervised setting. For many machine learning tasks, unlabeled data is plentiful, and labeled data is scarce. Semi-supervised learning aims to improve the accuracy of a model trained with supervision by exploiting information in unlabeled data [55]. Following [61], we can formally

define the labeled set as $\mathscr{D}_L = ((x_i, y_i))_{i=1}^{l}$ and the unlabeled set as $\mathscr{D}_U = (x_i)_{i=l+1}^{l+u}$, where l and u are the numbers of labeled and unlabeled samples, respectively. There is a great deal of methods with the goal of improving a model trained on $\mathscr{D}_L$ based on data from $\mathscr{D}_U$. These methods fall under the *Inductive* category in the taxonomy of self-supervised methods, shown in Figure 2.1. Whereas inductive methods yield a model that can be used to make predictions on unseen data, the transductive methods produce only the predictions directly.

One of the simplest approaches to semi-supervised learning, and also the most relevant approach for our work, is inductive learning based on wrapper methods. As described in [61], wrapper methods rely on a model trained with supervision on labeled data $\mathscr{D}_L$ to generate *pseudo-labels* on the unlabeled data $\mathscr{D}_U$. The model can then be retrained using a purely supervised approach, using a combination of the labeled and pseudo-labeled data, unaware of the distinction between the two, to attain the updated inductive model. The pseudo-labeling can be done with one or more supervised base models, and the process can be iteratively repeated to label more data and further improve the model. This basic approach can be extended in many ways. More details on wrapper methods are available in a comprehensive review in [62].

2.2 Deep learning

To improve the performance of conventional machine learning methods, early efforts in pattern recognition and machine learning fields were devoted to the design of novel and effective discriminant features [20]. Based on human ingenuity and often inspired by the human visual cortex, researchers designed a plethora of so-called *handcrafted* features. However, handcrafted features require considerable domain expertise and careful engineering to transform raw data to suitable representation [63]. Moreover, these methods are frequently designed for a specific task, and the ability to use them across different tasks and domains is very limited [20].

A novel approach, designed to mimic human neural systems with the goal of processing data in their raw form, was discovered independently by several research groups in the 1970s and 1980s [63]. The approach was based on Artificial Neural Networks and the Backpropagation algorithm [64, 65]. Despite its potential and success in some fields, this approach was largely dismissed in the 1990s due to feasibility concerns. Based on the concept of artificial neural networks, some of the main principles of Deep Learning (DL) were proposed in 2006 in a breakthrough work introducing the Deep Belief Network [66]. This was the basis for a new surge of development in the field. In general, deep learning methods are representation learning methods that start with the raw inputs (such as image pixel values) and obtain multi-level representations by simple but non-linear modules that each transform the representation at one level to a slightly more abstract representation at a higher level, facilitating the learning of very complex functions [63]. As reviewed in [56], the monumental success of deep learning methods

was boosted by the introduction of very large datasets, new training strategies (including new optimization and regularization algorithms), breakthroughs in computing power (GPU, NPU, TPU), the design of deep architectures, and transfer learning techniques, all based on the ability to learn from raw data.

We continue by introducing the basics of Artificial Neural Networks and the widely adopted Convolutional Neural Network. We conclude this section with a review of transfer learning, as it is one of the most significant deep learning concepts in our field of work.

2.2.1 Artificial Neural Network

The basic building block of the Artificial Neural Network (ANN) is the neuron. The neuron takes a set of inputs x_i and applies a set of weights W and a constant bias term b to form a linear function that approximates the input data [56]. The central idea of ANNs is to learn parameters W and b automatically based on the input data.

$$z = \left(\sum_{i=1}^{n} w_i x_i \right) + b \tag{2.1}$$

The output of the neuron is denoted as z and calculated from the inputs, weights, and bias according to Equation 2.1. Output z is used to calculate the activation $a = \sigma(z)$, where σ is the activation function. The primary purpose of the activation function is to introduce non-linearity, allowing the modeling of complex input-output relationships. Some common options for activation functions are the hyperbolic tangent function (*tanh*) and the *sigmoid* function. The predominant option in the deep learning field, however, is the Rectified Linear Unit (ReLU) [67], simply defined as $\sigma(z) = \max(0, z)$, as it is computationally efficient and it helps in the learning process of deep networks that often suffer from the so-called *vanishing gradient* problem [68, 69]. Figure 2.2 depicts the described processing steps in the neural network architecture.

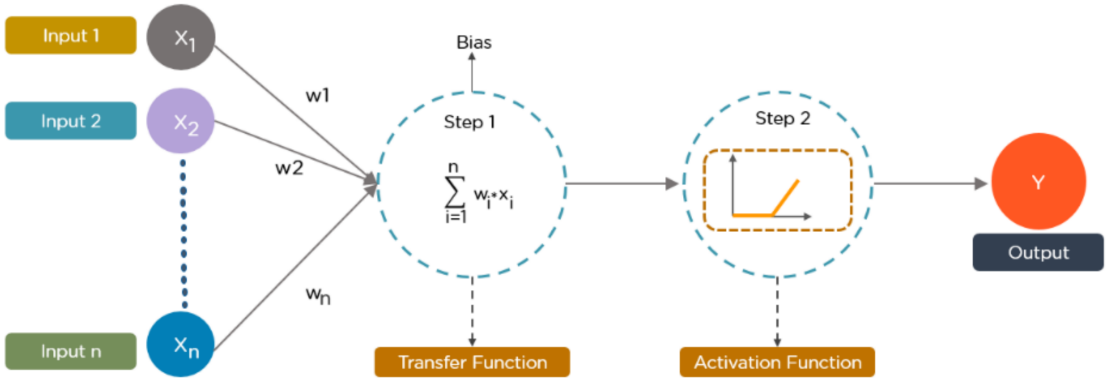

Figure 2.2: Data processing in neural network architecture. Reprinted from [70] under CC BY-SA 4.0 license.

It is common practice to group neurons into layers, to stack the layers successively, and to process them in a feedforward fashion. This forms the basis of the Multilayer Perceptron (MLP) architecture, depicted in Figure 2.3. MLP is referred to as the foundation architecture of deep neural networks (DNN) [71] and as the quintessential deep learning model [52]. A typical MLP consists of an input layer that takes the data, an output layer that provides the predictions, and a number of so-called *hidden layers* between the two, where the neurons are typically fully connected [71]. As formulated in [52], a feedforward network has the goal of approximating function $y = f(x)$ by learning the parameters θ and defining an optimal mapping $y = f(x;\theta)$. The basis of deep models is the multi-layered approach where function f is a chain function, e.g. $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$, where $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$ are the first, second, and third layer of the network, respectively. The number of layers d in $f^{(d)}$ defines the depth of the model. This type of network is called feedforward because the information flows from input x to output y through components that define f without any feedback connections, usually used in some other architecture types such as Recurrent Neural Networks (RNN) [52].

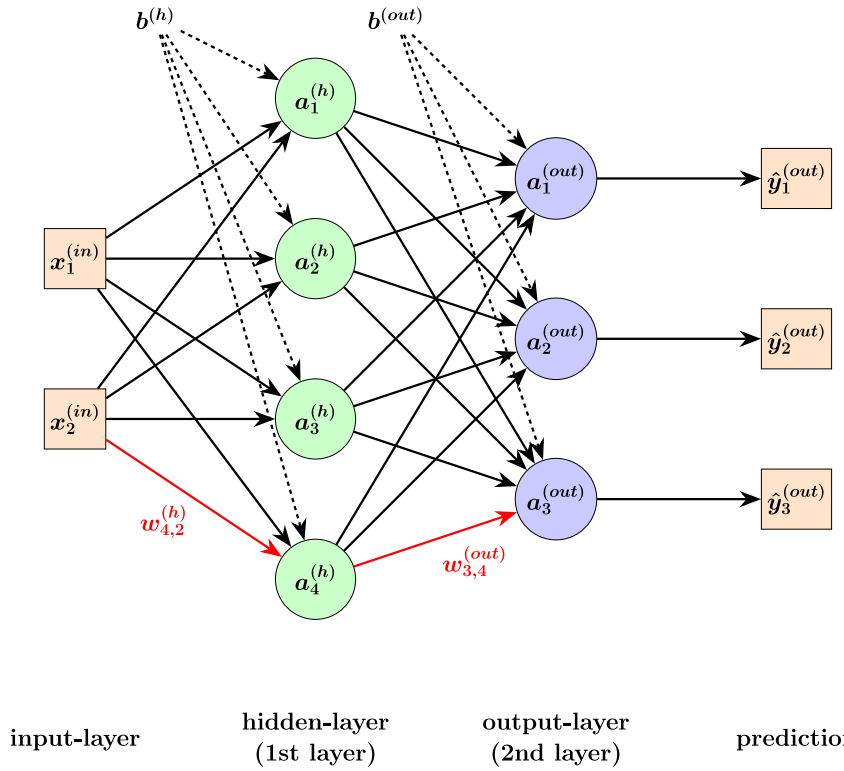

Figure 2.3: Multilayer Perceptron topology. Reprinted from [72] under CC BY-SA 4.0 license.

2.2.2 Convolutional Neural Network

Convolutional Neural Networks (CNN) are broadly defined as ANNs that have at least one layer where convolution operation is used instead of general matrix multiplication [52]. The idea of using deep multi-layered networks precedes CNN. The success of CNNs was built upon

three additional principles: local connections, shared weights, and pooling [63]. These design principles were motivated by visual neuroscience studies and the visual cortex [73].

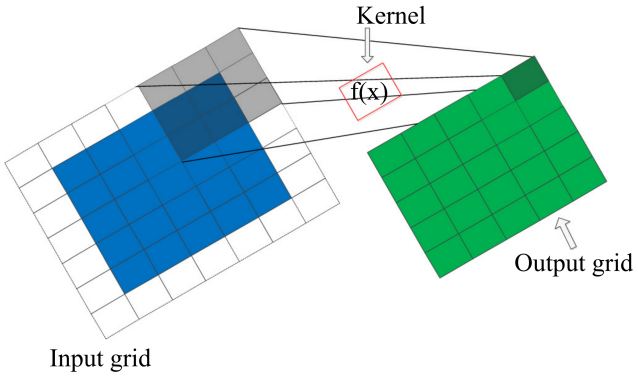

Figure 2.4: 2D convolutional operator. Reprinted from [74] with permission from Elsevier.

First, consider the convolutional operator, shown in Figure 2.4. The first argument in this operation is the input, the second argument is called the *kernel*, and the output is referred to as the *feature map*. Instead of connecting the input to the output with a fully connected network, CNNs use relatively small kernels with shared weights that are being shifted across the input to generate elements of the output feature map [73]. Since the weights are shared and the kernel is shifted, this design choice allows for the same local pattern to be detected in different input locations. The weight sharing also reduces the number of parameters, facilitating the design of deeper networks. The third design principle, referred to as pooling, is designed to reduce the dimensions of feature maps while aiming to merge semantically similar features [63]. The two most frequently utilized pooling operators are called *max pooling* and *average pooling*. Both these operators reduce the size of the feature map by splitting the input into a grid of local patches and by calculating a single output for each of the patches in the grid. To get a single value based on patch elements, simple maximum or average operations can be applied, according to the pooling operator type. The pooling operations support translation invariance, reduce computational complexity, and mitigate some of the issues associated with deep network training, such as overparameterization and overfitting.

The first popular CNN architecture, dubbed LeNet [73], was introduced in 1998 by LeCun et al. to classify handwritten digits. It was followed by some other popular architectures, such as LeNet-5 [76], AlexNet [77], and VGGNet [78]. While being very different with respect to the number of convolutional blocks, kernel sizes, activation functions, the pooling approach, and many other parameters, all these architectures follow a similar design pattern. The input image is processed by several convolutional blocks operating on different feature map sizes. Each block starts with one or more sequential convolutional layers paired with an activation function and ends with a pooling layer that reduces the feature map size. The feature maps, therefore, become progressively smaller and capture more and more abstract features. Finally,

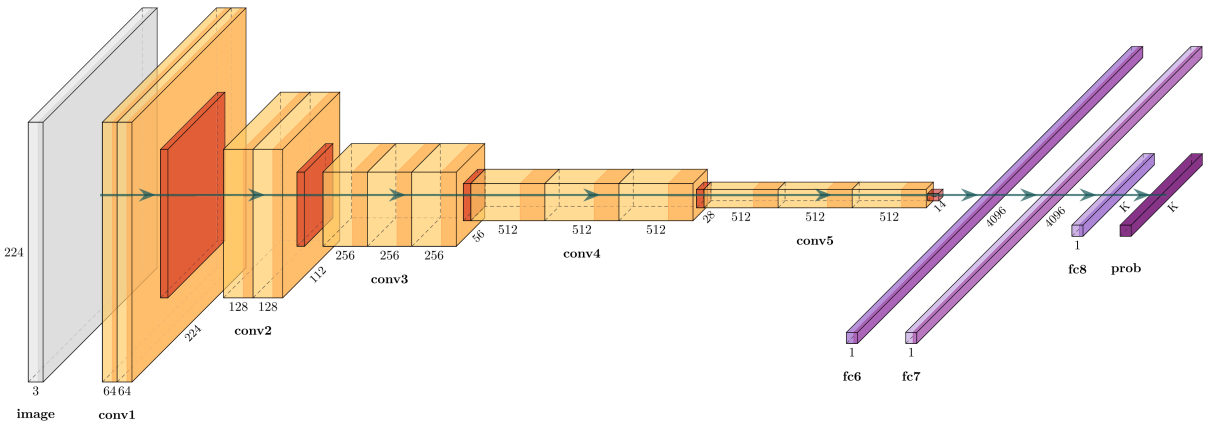

Figure 2.5: Architecture of the VGG-16 CNN. Reprinted from [75] under CC BY-NC-ND 4.0 license.

the flattened output of the last feature map is processed by a fully connected neural network to get the prediction vector. Figure 2.5 presents the architecture of the VGG-16 CNN. It starts with a 224×224 RGB image, followed by five convolutional blocks operating at different feature map sizes, and concluded by three fully connected layers, resulting in a vector containing probabilities for K classes.

More recent design concepts, such as global average pooling, identity mapping, residual connection, squeeze and expand module, and depthwise separable convolution, have enabled the design of many advanced architectures, such as GoogLeNet [79], ResNet [80], SqueezeNet [81], Xception [82], MobileNet [83], and MobileFaceNet [84]. Moreover, while CNNs are predominantly used for 2D data (e.g., images), they can also be used to process 1D data (e.g., temporal series) and 3D data (e.g., volumetric data). A relevant example of a CNN designed for temporal data processing is the Temporal Convolutional Network (TCN) [85].

2.2.3 Transfer learning

Deep CNNs were shown to have the ability to learn from vast amounts of data and perform complex tasks. One prominent example is the ImageNet Challenge [86], where models are trained using more than one million images to classify samples into 1,000 classes. However, training deep networks from scratch (i.e., starting with randomly initialized weights) is not trivial. Due to their depth and a large number of learnable parameters, deep networks require large amounts of training data and computationally-complex, time-consuming training procedures. A popular technique developed for the mitigation of this problem, called *transfer learning*, is based on knowledge transfer across domains, tasks, or distributions.

A common occurrence is that the target dataset $\mathscr{D}_T$ has a much lower amount of samples than some other available dataset $\mathscr{D}_P$. Although the dataset $\mathscr{D}_P$ might be sampled from a different distribution than $\mathscr{D}_T$, the model can learn some transferable information, such as how to handle

geometric and lighting changes or how to detect edges and blobs [52]. This can be useful across domains, so pretraining a model on $\mathscr{D}_P$ and then finetuning it on $\mathscr{D}_T$ can improve its generalization capabilities. Moreover, pretrained model weights can be found online for a large number of popular CNN architectures and some common pretraining datasets, such as ImageNet [87] and VGGFace [88]. This allows researchers to finetune models on the target data $\mathscr{D}_T$ much quicker, even with very limited computational resources. A popular pretraining dataset in the facial age estimation field is the IMDB-WIKI dataset [89].

2.3 Estimation methods

After introducing the most relevant machine learning methods and the basic deep learning architectures and concepts, we continue by formally defining relevant estimation methods. Once more, we narrow our scope and focus on estimation methods relevant to our work. Namely, deep learning methods based on the supervised learning scheme. The two predominant estimation method categories from this field are classification and regression. Advanced methods often blur the line between those two categories. In this section, we revisit the basic formulation introduced in Section 2.1.1 and use it to define the most relevant methods.

Formally, in a dataset $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i \in 1,\ldots,N}$ with N samples, let $x_i \in \mathbb{R}^{h \times w \times c}$ denote i-th image sample where h, w, and c are the image height, width, and number of channels, respectively. Assume that $y_i \in \{0, 1, \ldots, K\}$ is the corresponding label with a range from 0 to K. Parameters of the deep learning model are denoted as θ while $z_i \in \mathbb{R}^K$ represents the model output feature vector, where z_i is obtained as $z_i = f(x_i; \theta)$.

As the focus of our work is age estimation, let us consider a special case where y_i can be treated as a category from a limited set of K ordered classes or a discrete number limited to the $[0, K]$ range. As reviewed in [20], it should be taken into account that the order relationship of age labels is intrinsically defined due to the nature of the age estimation problem. Treating age labels as unrelated categorical values is a flawed approach. For example, if the ground truth age is 30, predicting 50 should be considered more erroneous than predicting 31.

2.3.1 Classification

Classification refers to the category of algorithms that aim to map the inputs $\mathbf{x}_i$ to a finite set of label categories $Y \in \{0, 1, \ldots, K\}$. The most widely used approach in the deep learning field relies on the Softmax function and Cross Entropy Loss. This method is commonly referred to as the Softmax method. The Softmax function is applied to the model output z_i to obtain the estimated probability distribution $\hat{p}_i \in \mathbb{R}^K$. The $\hat{p}_{i,j}$ is calculated as:

$$\hat{p}_{i,j} = \frac{e^{z_{i,j}}}{\sum_{k=0}^{K} e^{z_{i,k}}}, \tag{2.2}$$

for $j \in \{0, 1, \ldots, K\}$. To optimize the model parameters θ, we can utilize Cross Entropy Loss defined as:

$$L_S = \frac{1}{N} \sum_{i=1}^{N} -y_i \log \hat{p}_{i,y_i} \tag{2.3}$$

The prediction, belonging to a finite set of categories $Y \in \{0, 1, \ldots, K\}$, is then calculated as:

$$\hat{y}_i = \underset{j}{\operatorname{argmax}}(\hat{p}_{i,j}) \tag{2.4}$$

2.3.2 Regression

Regression is the second widely adopted category of estimation algorithms. Regression algorithms aim to map the inputs $\mathbf{x}_i$ to a continuous label value, meaning that the model needs to output a single value $\hat{y}_i \in \mathbb{R}^1$. To accommodate the previously formalized setting, we can calculate the prediction $\hat{y}_i$ based on the model output vector z_i as

$$\hat{y}_i = \sum_{j=0}^{K} j * \hat{p}_{i,j}, \tag{2.5}$$

where $\hat{p}_{i,j}$ once again denotes the estimated Softmax distribution from Equation 2.2. To optimize the model parameters θ, we can adopt the widely used Mean Square Error Loss (MSE), defined as:

$$L_R = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.6}$$

The most common alternative to the MSE loss for regression is Mean Absolute Error Loss (MAE). MSE loss discourages outlier predictions with huge errors, as the squaring term magnifies the error. This can be advantageous, but can also put too much attention on outliers and potentially faulty labels. On the other hand, all the errors calculated by MAE are weighted equally. MAE is defined as:

$$L_A = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{2.7}$$

2.3.3 Advanced estimation methods

Regression and classification algorithms can be combined and modified in many ways, resulting in a multitude of methods originating in many different machine learning fields. We conclude

this section with four advanced methods that originated, or are proven to excel, in the age estimation field.

Deep Expectation

Similar to the Softmax method, the Deep Expectation (DEX) method [89] uses the Softmax function from Equation 2.2 and Cross Entropy Loss defined by Equation 2.3 to optimize model parameters θ. However, contrary to the Softmax method, estimated probability distributions are used at inference time to calculate a continuous value as a weighted average defined by Equation 2.5. This method uses classification-based learning and regression-based inference.

Mean-Variance

The second advanced method extends the DEX idea introduced in [89] by enhancing the basic Softmax Loss with two additional components: Mean Loss and Variance Loss. The Mean-Variance Loss [90] is defined as:

$$L_{MV} = \frac{1}{N}\sum_{i=1}^{N}\left(-y_i \log \hat{p}_{i,y_i} + \frac{\lambda_1}{2}(y_i - \hat{y}_i)^2 + \lambda_2 \sum_{j=0}^{K} p_{i,j} * (j - \hat{y}_i)^2\right), \tag{2.8}$$

where factor λ_1 controls the contribution of the Mean Loss component, while λ_2 controls the contribution of the Variance Loss. This advanced method combines classification and regression-based components with an additional distribution-oriented variance component at training time. The predictions are calculated according to Equation 2.5.

Deep Label Distribution Learning

The third advanced method deemed relevant for our work is DLDL-v2 [91]. This label-distribution-based method trains the model parameters θ by utilizing a loss function defined as:

$$L_{LD} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=0}^{K} -p_{i,j} \log \hat{p}_{i.j} + \lambda \mid y_i - \hat{y}_i \mid\right), \tag{2.9}$$

where p_i denotes an approximated ground truth label distribution. The aforementioned distribution is calculated based on the ground truth age label y_i as:

$$p_{i,j} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(j-y_i)^2}{2\sigma^2}} \tag{2.10}$$

for $j \in \{0, 1, \ldots, K\}$. The rationale for representing age labels as distributions lies in the ordered nature of age labels, as the degree to which sample x_i is described by labels neighboring to y_i (e.g., $y_i - 1$ or $y_i + 1$) should be higher than for some more distant labels (e.g., $y_i - 10$ or $y_i + 10$). This advanced method combines the distribution-based Kullback-Leibler Divergence Loss and regression-based MAE Loss. The parameter λ controls the balance between the loss components. The method predicts continuous values $\hat{y}_i \in \mathbb{R}^1$, once again calculated according to Equation 2.5.

Ordinal ranking

The final method in this review is based on the ordinal ranking approach. The previously introduced methods rely on regression, classification, or an advanced combination of the two approaches. The differences between those methods can be boiled down to forming different label representations and defining more or less complex loss functions to accommodate different optimization objectives. On the other hand, the ordinal ranking methods rely on a framework based on aggregation of multiple binary classifiers, as proposed in [92]. The Cross Entropy Loss, defined in Equation 2.3 and also used in the regular multi-class classification approach, can be used to optimize each of the binary classifiers in this framework. In fact, any type of binary classifier can be employed. The task of mapping inputs x_i to a finite set of label categories $Y \in \{0, 1, \ldots, K\}$ is approached by a set of K binary classifiers, where k-th classifier is optimized to simply estimate if the label associated with the input x_i is higher than k or not. Given a ground truth label $y_i \in \{0, 1, \ldots, K\}$, the first y_i classifiers are trained to produce a positive answer, while the rest are trained to produce a negative answer. According to this, the prediction $\hat{y}_i$ is equal to the rank $r(x_i)$, which can be calculated as:

$$\hat{y}_i = r(x_i) = \sum_{k=0}^{K-1} [\![O_k(x_i > 0)]\!] \tag{2.11}$$

Here, O_k is the output of the k-th ordered binary classifier, and $[\![\cdot]\!]$ is 1 if the enclosed condition is true, and 0 otherwise. This formulation is designed to divide a complex task into a series of simpler sub-tasks and also to take into account the relative order of the labels. This is very much appropriate for the age estimation task, as the nature of the task entails ordered labels.

2.4 Significance of data

We conclude this chapter by briefly reviewing the role of data in machine learning. Advancements in learning algorithms and model architectures, along with large amounts of available data and computing infrastructure, have enabled researchers to design methods that surpass human performance on difficult tasks such as image classification [93] and face recognition

[94]. The main remaining barrier to solving many similar tasks is the lack of sufficient labeled data. While transfer learning is frequently utilized to mitigate this problem and achieve state-of-the-art results, training with small numbers of task-specific samples can still result in domain overfitting, questionable generalization capabilities, and unsatisfying performance in an unconstrained environment. For these reasons, the acquisition of a suitable training set is considered to be the most critical step in the development of a machine learning model [42].

As discussed in Section 2.1.1, the supervised learning approach requires large amounts of labeled data. Collecting labeled data can be a laborious process, requiring manual annotation and a priori knowledge or even domain expertise. Although using manual annotation and expert knowledge leads to great results, this approach is often not scalable. In the case of facial data collection, privacy concerns related to biometric data storage and processing also need to be taken into account. To support generalization in unconstrained conditions, facial datasets need to capture variability with respect to age, gender, ethnicity, head pose, recording environment, lighting conditions, occlusion level, recording quality, and so forth. Due to all these challenges, the collection of a *perfect* dataset is often deemed unfeasible and public datasets frequently suffer from unbalanced sample distributions and data disparity [42].

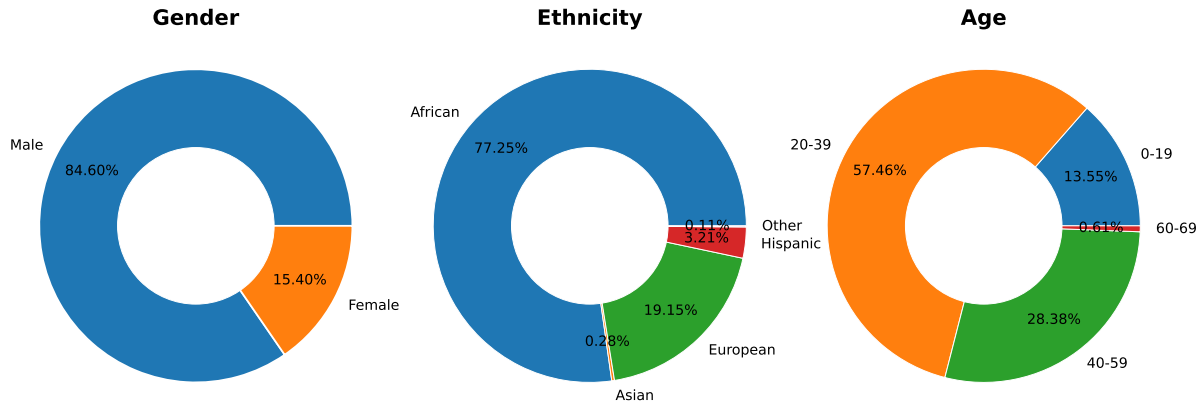

Figure 2.6: MORPH-II benchmark dataset [95] sample distribution across gender, ethnicity, and age.

As most research groups don't have the resources to carry out complex data collection projects, public data can be considered the backbone of the machine learning research field. Public benchmarks allow researchers to design, evaluate, and fairly compare methods on common data, thus stimulating the growth of related research fields [20]. A benchmark protocol consists of benchmark data, performance metrics, and the evaluation protocol determining how to calculate the defined metrics on the given data. The benchmark dataset (i.e., test set) must be independent of the training data to evaluate generalization capabilities on unseen data. A good benchmark dataset is also a well-balanced representative of the target real-world environment, sufficiently covering its variability. Using loosely defined benchmark protocols with inappropriate test data can lead to faulty or biased conclusions. To illustrate the problem of

imbalanced data in our field, Figure 2.6 shows gender, ethnicity, and age distributions for the most frequently used age estimation benchmark dataset. This benchmark protocol proposes the use of different folds from the dataset for both training and testing. It is reasonable to conclude that even a high-performing model developed and evaluated on this data can have very bad generalization to real-world data. Data imbalance and quality issues related to training data can, to some extent, be mitigated by contemporary training techniques. However, in the case of test sets used for benchmarking, it is of utmost importance to use high-quality data and to make it publicly available, as this enables researchers to draw valid and verifiable conclusions based on transparently comparable results.

Chapter 3

Automatic face analysis

Research on automatic face analysis includes many topics, such as face recognition [14], age estimation, [3], gender classification [17], ethnicity identification [18], expression recognition [15], and eye gaze analysis [16]. While humans tend to analyze faces with a holistic approach, where identity, age, gender, and ethnicity are simultaneously taken into account, researchers usually focus on designing algorithms that tackle each of those tasks separately. While single-task algorithms show superior performance in some cases, studies also suggest a strong interaction between facial attributes. Some multi-task approaches combine age and gender [96], age and expressions [97], age and ethnicity [22], and age with both ethnicity and gender [19] or both expressions and gender [98]. Studies also indicate that learning age jointly with gender, ethnicity, or expressions is a more challenging task than learning age independently [43].

The literature introduces various approaches for each of the aforementioned face analysis tasks. However, most of them share a very consistent general face analysis framework. We implement this face analysis framework in our own research and focus on single-task facial age estimation, following the motivation presented in Section 1.1. Details on the general face analysis framework, followed by the specifics of the facial age estimation framework, are given in the following sections. This chapter builds upon our previous work on automatic image-based face analysis from [99].

3.1 General face analysis framework

The general face analysis framework can be described as a pipeline consisting of four consecutive processing steps. Since this generic framework is utilized by a vast majority of published face analysis methods, similar processing patterns were already identified and described in some relevant face analysis reviews [2, 3, 20]. As in many other image analysis frameworks, the first step is object (e.g., face) detection, followed by preprocessing, feature extraction, and attribute prediction steps. The pipeline is presented in Figure 3.1.

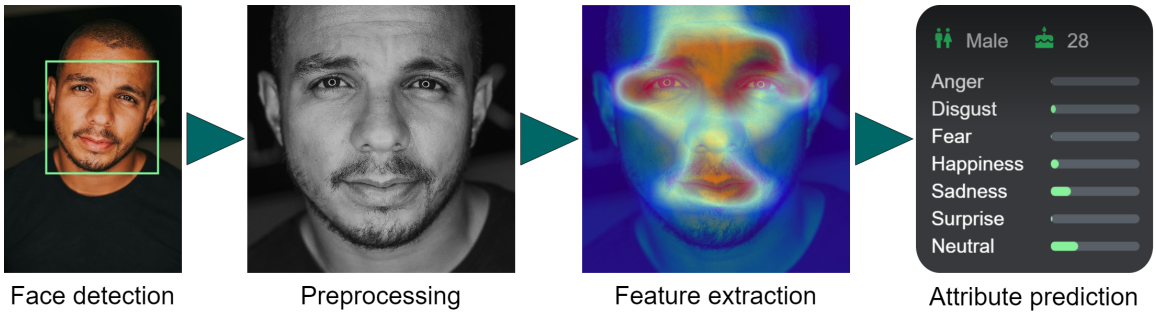

Figure 3.1: The four steps of the general face analysis pipeline.

3.1.1 Face detection

The face analysis pipeline starts with the face detection step. This step results only in very basic face location and scale information, usually provided via a facial bounding box. As this step seems fairly trivial, it is often not highlighted. Nevertheless, different versions of face detectors can be trained with differently defined bounding boxes, and can result in inconsistent detections, thus introducing bias and propagating error to the rest of the pipeline. To some extent, this can be mitigated using the preprocessing techniques described in the following section. However, in case of false positive or false negative detections, the impact on performance can not be avoided.

The most widely used face detection systems are based on the work of Viola and Jones [100], and, more recently, deformable parts models [101], cascaded CNN detectors [102], single shot detectors [103], region proposal CNNs [104], and detectors based on Feature Pyramid Networks [105]. Of special interest is the RetinaFace detector [106], which simultaneously predicts the face bounding box and five facial landmark points. This additional output can be very useful in later processing steps. More details on face detection methods can be found in [107, 108, 109].

3.1.2 Preprocessing

Depending on the implemented face analysis method, the preprocessing step can be as trivial as image cropping based on the detected facial bounding box. Additional minimal preprocessing techniques include resizing the face crop image and color space conversions. However, studies show that more advanced preprocessing steps can significantly improve the final performance [2]. The main goal of preprocessing is to provide consistent inputs and capture important information for the downstream pipeline steps (i.e., feature extraction and attribute prediction).

The primary sources of input inconsistencies are face detector inaccuracies and variations in head pose angles. Several pose normalization and image alignment techniques have been proposed to improve input consistency. The most basic method relies on face detection confidence. The input image is rotated by a small angle multiple times, and the version with the highest detection confidence is used. Although computationally expensive, this simple method does not introduce any new components to the pipeline, as the existing face detection system is reused,

and in some cases results with satisfying performance [89], [110]. The more frequently used method does introduce an additional component to the pipeline, as it relies on face landmark detection. Systematic overviews of face landmark detection methods are given in [111, 112, 113]. Given a set of detected landmark points, in-plane rotation and scaling can be performed based on eye points [114], [115]. A fitted 3D face model can be used to align faces along multiple axes [116]. The most frequently adopted landmark-based method aligns the face by computing an affine transformation which puts the eyes, nose, and mouth into canonical coordinates [20], as shown in Figure 3.2.

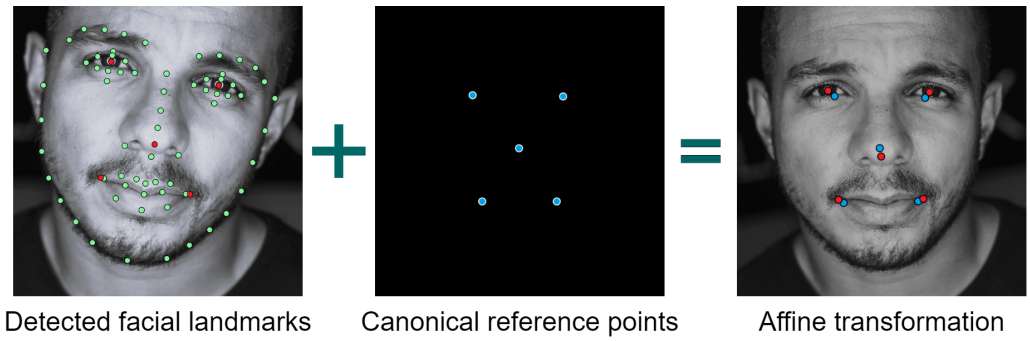

Figure 3.2: Affine transformation aligning the eyes, nose, and mouth into canonical coordinates.

Another source of input inconsistencies are variations related to lighting and imaging conditions. Basic preprocessing often includes intensity standardization or normalization. To deal with low contrast and brightness problems, techniques such as Difference of Gaussians [117] and Contrast Limited Adaptive Histogram Equalization (CLAHE) [118] are sometimes incorporated.

3.1.3 Feature extraction

It is well known that any model is only as good as the data it works with. Accordingly, feature extraction is the most emphasized step in many publications. The goal of the feature extraction step is to extract informative, discriminative, and compact feature vectors that can be efficiently used in the final attribute prediction step.

Before the advent of learnable feature extractors, a lot of focus was put on the design of handcrafted features. Local Binary Patterns (LBP) [119], Biologically Inspired Features (BIF) [120], Histogram of Oriented Gradients (HOG) [121], and Speeded up Robust Features (SURF) [122] are some of the most popular options in the face analysis field. Combinations of different types of features can be used, but this can lead to impractical feature sizes. To alleviate this problem, dimensionality reduction methods such as Principal Component Analysis (PCA) [57] and Linear Discriminant Analysis (LDA) [123] are often used. Despite all the efforts put into the development of handcrafted feature representations, their expressiveness remained quite limited and they are deemed adequate mostly for small datasets and constrained frontal faces.

More recently, Deep Learning has become the dominant paradigm in the face analysis field, as it is a consolidated conviction that it offers overwhelming superiority over dated methods based on handcrafted features in general [20]. Deep Learning feature extractors are optimized to learn discriminative representations directly from raw pixels [2] and often based on millions of images, as opposed to feature extractors that were designed manually, and based on human ingenuity and intuition. More details on handcrafted and learning-based feature extractors are given in Section 4.3.

3.1.4 Attribute prediction

The goal of the attribute prediction step is to infer estimates related to the target face analysis task based on feature vectors extracted in the previous step. Depending on the face analysis task, the expected output can be a class from a predefined set (e.g., gender or ethnicity estimation) or a continuous value (e.g., age estimation). The prevalent attribute prediction algorithms for those two types of estimation are classification and regression, respectively. Classification can be binary (e.g., basic gender estimation) or multi-class (e.g., ethnicity estimation). Interestingly, some tasks can be formulated both as classification and regression problems. For example, the age value can be estimated directly with the regression method or indirectly with the classification method, where the classes are defined to cover some predefined age range (e.g., 101 classes for ages from 0 to 100). Moreover, there are some more advanced estimation methods, often combining classification and regression. Some more details on advanced approaches for attribute prediction are given in Section 2.3.3.

The classic machine learning approaches, where feature extraction is primarily based on handcrafted features, often perform attribute prediction utilizing some of the classic regression and classification methods. Support Vector Machines (SVM) [124], Support Vector Regressors (SVR) [125], Random Forests [126], K-Nearest Neighbors (KNN) [127] or boosting algorithms such as AdaBoost [128] are frequent choices from an otherwise extensive list of options.

Contemporary face analysis methods are primarily based on the Deep Learning paradigm and Convolutional Neural Networks. One of the main advantages of this approach is the ability to perform a so-called *end-to-end* optimization where feature extraction and attribute prediction steps are optimized jointly. Deep learning methods can be designed both as classification and regression methods, but the deep learning paradigm also offers more flexibility in the design of the estimation task and the optimization method. This enables many advanced attribute prediction approaches, often built upon work from other machine learning fields. A review of the advanced estimation approaches for age estimation is given in Section 4.2.3.

Due to the prevalence of end-to-end learning in contemporary methods, the feature extraction and attribute prediction steps are often not distinguished as separate steps. Recent works usually emphasize only the preprocessing details and the design of the deep learning method.

3.2 Age estimation framework

The rudimentary facial age estimation framework is very well aligned with the general face analysis framework described in the previous section. A pattern with the same four consecutive steps, including face detection, preprocessing, feature extraction, and attribute prediction, is identifiable in most of the reviewed age estimation methods. Specifics of the age estimation framework covered in this section are mainly related to the attribute prediction step. The main focus of our work is chronological age estimation. Related research fields include apparent age estimation and age group classification. For completeness, we briefly discuss all three attribute prediction approaches, with a special focus on the related age estimation metrics.

3.2.1 Chronological age estimation

Chronological age estimation is the process of estimating the number of years elapsed from one's birth to a certain point in life [4]. It is the most challenging type of age estimation. Compared to the apparent age estimation, it is more difficult due to the personalized and stochastic nature of the aging process, and the influence of intrinsic and extrinsic factors on facial appearance, as elaborated in Section 1.1.2. Moreover, while chronological age estimation requires the estimation of a continuous age value or a precise age category (e.g., 100 categories), the goal of the age group estimation is a prediction of the age category from a more limited set. Intuitively, it is easier to estimate if the subject is a child or an adult than it is to estimate its exact age. Chronological age estimation is also a type of age estimation that has the greatest coverage in literature, and is a requirement for most of the application fields described in Section 1.1.1. Along with the term chronological, this type of estimation is also referred to as real or actual age estimation.

Formally, in a set with N samples, let $x_i \in \mathbb{R}^{h \times w \times c}$ denote i-th facial image sample where h, w, and c are the image height, width, and number of channels, respectively. For the chronological age estimation task, each image sample in such a set requires a corresponding ground truth chronological age label $y_i \in \{0, 1, \ldots, K\}$, assuming an age range from 0 to K. The age is estimated either as a continuous value $\hat{y}_i \in \mathbb{R}$ or as a precise age category $\hat{y}_i \in \{0, 1, \ldots, K\}$, depending on the type of the estimation algorithm.

To measure how far estimate $\hat{y}_i$ as from ground truth y_i, regardless of whether it is too high or too low, we make use of the Absolute Error (AE) metric, defined as:

$$AE_i = |y_i - \hat{y}_i| \tag{3.1}$$

The most frequently used metric for chronological age estimation is Mean Absolute Error

(*MAE*). It is defined as the average of absolute errors calculated over N test samples:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} AE_i = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{3.2}$$

Good performance is indicated by a low *MAE* score. *MAE* is a good choice of a metric in cases where the set has missing samples [2], meaning that the set doesn't have a dense distribution covering the complete age range. This formulation does not provide performance for specific ages, but only general performance. To further analyze performance with respect to specific ground truth age, a modification of the *MAE* metric can be defined. Similarly to [4], we denote one specific age with k and calculate *MAE* over all samples that have the ground truth age label y_i equal to k. Given that n_k denotes the total number of samples where y_i is equal to k, MAE_k is defined as:

$$MAE_k = \frac{1}{n_k}\sum_{i=1}^{n_k} |k - \hat{y}_i| \tag{3.3}$$

Cumulative Score (*CS*) is another metric appropriate for measuring overall age estimation performance. Given that $n_{AE \leq T}$ is the number of samples where the age estimation absolute error (*AE*) is no higher than a predefined threshold of T years, and N is the total number of samples in the set, *CS* is defined as:

$$CS(T) = \frac{n_{AE \leq T}}{N} \times 100\% \tag{3.4}$$

Good performance is indicated by a high *CS* score. This metric is most suitable when the set has samples for every age [2] (i.e., when the set has a dense distribution). As *CS* is a function of threshold T, the results for this metric can be given as curves changing with respect to T.

3.2.2 Apparent age estimation

Apparent age estimation is an estimation of age perceived by other humans. When estimating chronological age, the model needs to learn to compensate for various intrinsic and extrinsic factors that influence facial appearance. On the other hand, the goal of apparent age estimation is to estimate how old the subject in the image appears to be, without trying to estimate its real age. This makes it a considerably easier task. According to the experiments in [30], apparent age correlates better with the face image, as *MAE* was reduced by more than 25% when chronological age labels were replaced by apparent age labels. Figure 3.3 depicts differences between apparent and chronological age labels. Apparent age estimation is also referred to as perceived age estimation.

Compared to chronological age, apparent age datasets also have different labeling requirements. For a set with N samples, formally defined in the previous section, a set of human votes

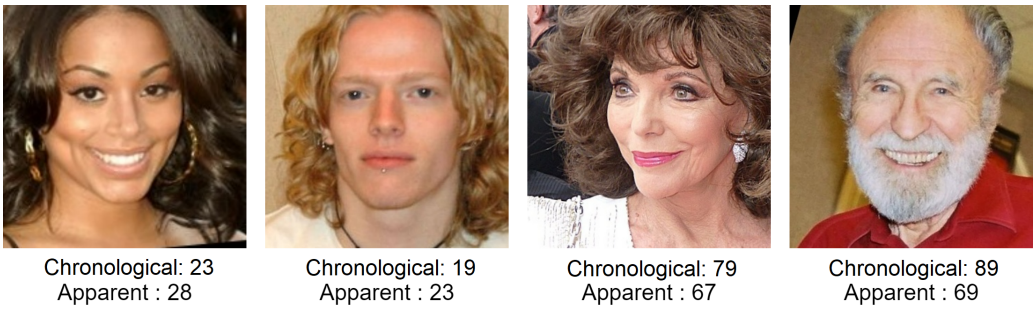

Figure 3.3: Differences in the chronological and apparent age labels in the Appa-Real dataset [30].

(i.e., age estimates) is required for each sample x_i. The votes are used to calculate mean age μ_i and standard deviation σ_i for each sample. Even though a single vote can be used as the apparent age label, more votes result in a more reliable ground truth, reflecting the so-called *wisdom of the crowd*. μ_i and σ_i are used as the ground truth for the i-th sample. While apparent age *MAE* can be calculated by substituting y_i with μ_i in Equation 3.2, the prevalent metric in the apparent age estimation field is the ε-error [29], formally defined as:

$$\varepsilon_i = 1 - e^{-\frac{(\hat{y_i}-\mu_i)^2}{2\sigma_i^2}} \tag{3.5}$$

The value of the ε-error can be between 0 and 1, with good performance being indicated by a low score. This metric also considers the difficulty of each of the samples. In particular, the samples with high standard deviation contribute less to the error score. A high standard deviation shows that there is a high level of inconsistency in the collected human votes, indicating that it was difficult for the voters to estimate age correctly. Such samples will likely be difficult for the estimation models as well. ε-error is also referred to as ε-score or Normal Score.

3.2.3 Age group estimation

Subjects whose real age falls within a predefined age range are said to belong to the same age group [4]. The limits and the number of age groups can be defined arbitrarily. The simplest example is based on just two age groups: minors and adults. The literature usually proposes a larger number of age groups (e.g., 7 [129] or 8 [130]) that are asymmetrical in size and cover some distinguishable stages of life. While age group estimation is linked to the chronological age in [4], labels for age groups are often estimated manually [129, 130]. This means that, in some cases, the age group labels are actually apparent age group labels. Therefore, age group estimation is a special category of age estimation that can be related either to apparent age or to chronological age estimation, depending on label type.

The most common metric for age group estimation is classification accuracy (*ACC*), defined

as the ratio between the number of correctly classified samples and the total number of samples:

$$ACC = \frac{n}{N} \times 100\% \tag{3.6}$$

Here, n represents the number of samples that are categorized in the correct age group, while N once again denotes the total number of samples in the set. In addition to this, literature related to age group estimation, as well as to many other classification-based fields, defines the 1-off metric as:

$$1-\text{off} = \frac{n_a}{N} \times 100\% \tag{3.7}$$

For this metric, n_a represents the number of samples that are categorized into the correct age group or an adjacent (1-off) group. Hence, this metric is more tolerant and results in higher scores than the standard *ACC* metric. For both of these metrics, a higher score naturally indicates better performance.

Chapter 4

Related work

The generalized age estimation framework, discussed in the previous chapter, consists of face detection, preprocessing, feature extraction, and attribute prediction steps. While face detection and preprocessing are integral parts of the processing pipeline, these steps are usually not emphasized in the age estimation literature. Most of the reviewed work puts focus on the design of effective feature extractors and novel attribute prediction algorithms. To systematically review the related work, we first propose a taxonomy of age estimation system design choices. The choices related to the feature representation and estimation algorithm are, therefore, the two main factors in the proposed taxonomy. Due to recent advancements in the exploration of video-based methods, we add the input data type as another top-level dividing factor in the hierarchy. The proposed taxonomy is presented in Figure 4.1.

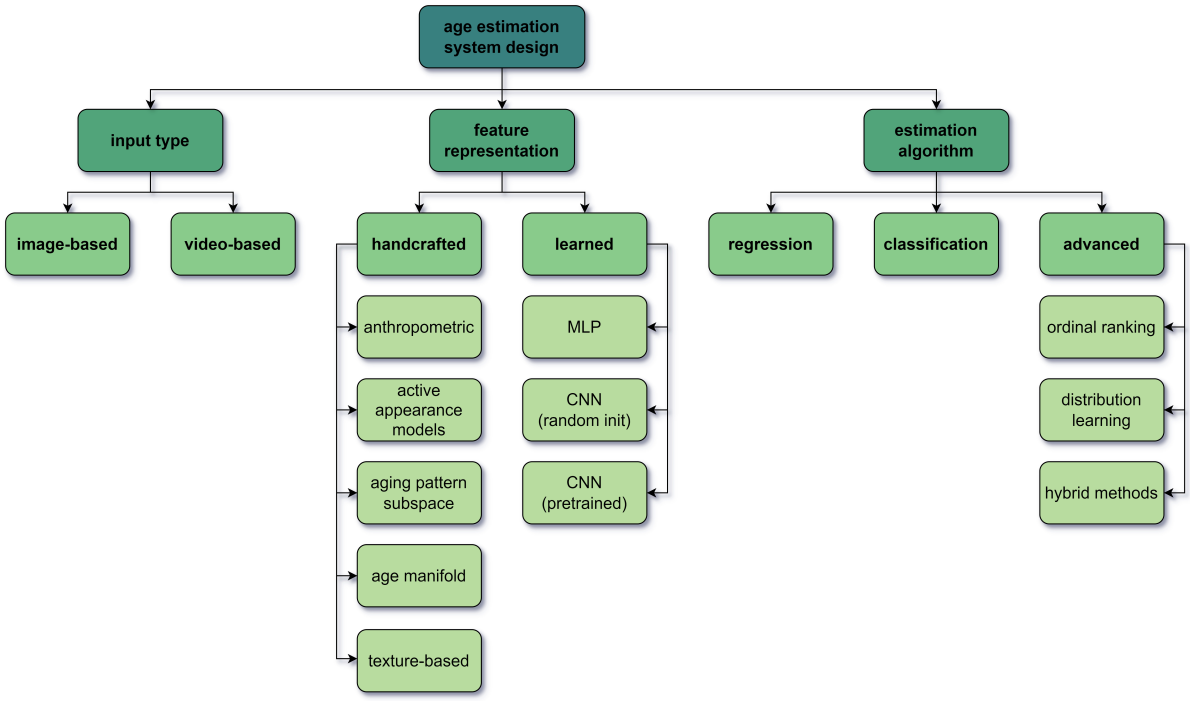

Figure 4.1: The taxonomy of age estimation system design choices.

The importance of data in the machine learning field is briefly discussed in Section 2.4. As data is a foundational component required for the design and evaluation of all age estimation methods, we start our review by introducing the most important age estimation datasets and benchmarks. Considering the proposed taxonomy, we continue by reviewing methods according to the used estimation algorithm. Finally, we review methods based on the feature representation approach, both for image-based and video-based methods.

4.1 Datasets and benchmarks

As discussed in Section 3.2, chronological age estimation is the task of estimating a subject's actual age, as opposed to apparent age estimation, which refers to the estimation of age as humans perceive it, based on physical appearance. A dataset can be annotated with chronological or apparent age or both. Whereas early research was primarily focused on chronological age estimation, apparent age estimation recently started to gain traction. Moreover, some datasets provide precise age labels, while others only provide age groups. Besides that, early work was mostly focused on age estimation in a controlled environment, while contemporary work shifted towards the more difficult task of age estimation in unconstrained in-the-wild conditions. The terms *in-the-wild* and *unconstrained* are used when data is collected without predefined limitations related to head pose, occlusions, background, lighting, image quality, and other relevant properties. This section introduces the most important image and video-based datasets, focusing on label type, collection procedure, dataset size, and sample distribution.

4.1.1 Image-based datasets

Early research on automatic facial biometric trait estimation was conducted on small manually collected image datasets, often having less than 100 samples [131, 132]. The introduction of larger datasets, such as FG-NET and MORPH, initiated an increase of interest in research related to age estimation, making it into an established research field over the past two decades [133]. The majority of datasets used in this field are image-based datasets. Samples taken from the most widely used datasets are presented in Figure 4.2, while Table 4.1 summarizes the main properties of the reviewed datasets.

Manually collected datasets

The majority of age estimation research is based on manually collected data. Manual data collection can ensure precise age labels and high-quality images. However, it often implies a controlled environment, limited data size, and constrained data distribution. The following are the most relevant examples of manually collected datasets used in the field.

The Face and Gesture Recognition Research Network Dataset (FG-NET) [35] was one of the first publicly available datasets for facial age estimation, initially introduced for studying facial appearance changes caused by aging. Over the years, it has supported a large volume of research from many fields, such as age-invariant face recognition, age progression, and age estimation, as reviewed in [133]. It is a cross-age dataset comprising 1,002 images of 82 subjects from different ethnic groups. This amounts to around 12 age-separated images per subject. To collect the dataset, subjects were asked to scan their personal photo collections, including childhood photos. The nature of this data collection process caused the images to be of varying quality, as it depended on the condition of the photographic paper, camera type, and scanning equipment. This dataset is considered an in-the-wild dataset since the images display variability in terms of resolution, clarity, illumination, head-pose orientation, facial expression, and facial occlusion (e.g., hats, eyewear, and facial hair). Although small in size, this manually collected dataset was a difficult challenge and an important steppingstone for the early age estimation research.

The Craniofacial Longitudinal Morphological Face Database (MORPH) [95] is another important building block in the contemporary facial age estimation research, created by the Face Aging Group at the University of North Carolina. The dataset was collected in a correctional facility over a period of 4 years. The images are collected in a controlled setting, and the image quality and resolution are rather poor. Its popular, publicly available subset (MORPH-II) consists of 55,134 images from 13,000 subjects. It provides annotations for age, gender, and race. Even though the images were collected in a highly controlled environment and the dataset has an unbalanced distribution of samples across gender (85% male), age (80% between 20 and 50 years, no children or elderly subjects), and race (77% African American), it increased the number of publicly available samples for age estimation research by a factor of 55 and made a great impact in the field. According to [28], MORPH-II is by far the most frequently used dataset for chronological age estimation.

The Appa-Real Dataset [30] is a more recent manually collected dataset, related to data from CLAP 2015 [29] and CLAP 2016 [134] challenges. The authors designed a data collection and labeling web application and utilized the Facebook API and the Amazon Mechanical Turk platform to get diversified data. The Appa-Real dataset consists of 7,591 unconstrained in-the-wild samples with age range from 0 to 95. Whereas the ChaLearn LAP datasets provided only apparent age labels, the Appa-Real dataset is extended with real (i.e., chronological) age labels, making it the only dataset that provides highly reliable labels for both real and apparent age. The authors utilized their platform to collect almost 300,000 apparent age votes, which amounts to approximately 38 votes per image. Having less than 8,000 samples, this dataset can be considered a small dataset in the era of deep learning. However, according to [20], it is the most challenging and well-rounded benchmark option.

The AgeDB Dataset [135], as the title of the reference paper claims, is a manually collected in-the-wild dataset. While the images and metadata originate from web sources, the data is collected manually. The Google Image Search* platform was used to collect images of famous people with information related to the exact age of the subject explicitly mentioned in the image caption. The dataset comprises 16,488 images from 568 subjects, covering an age range from 1 to 101 years. The dataset also provides gender and identity labels.

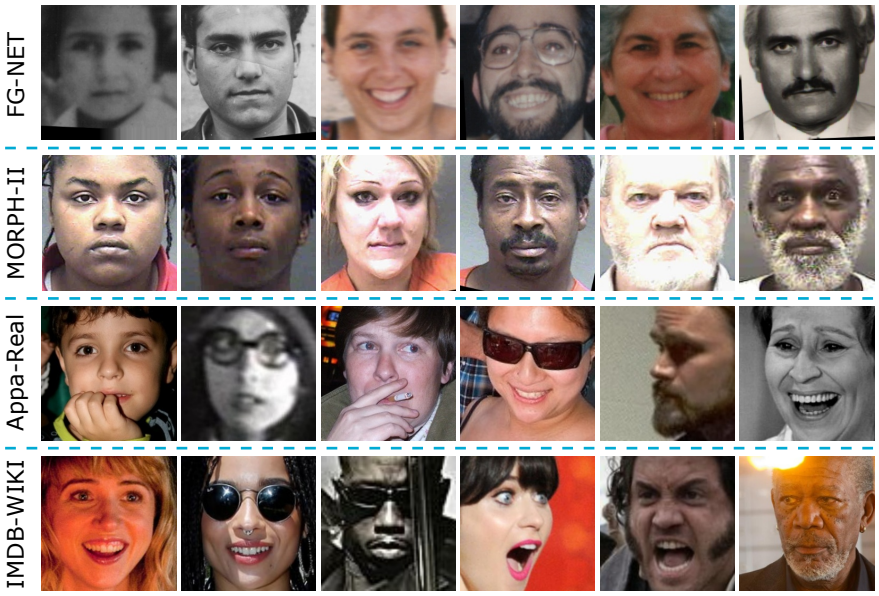

Figure 4.2: Aligned face crops based on images from the four most popular age estimation datasets. The FG-NET dataset consists of scanned album photos, the MORPH-II dataset is collected under controlled conditions in a correctional facility, Appa-Real is a crowdsourced dataset, while IMDB-WIKI contains web-scraped images of celebrities.

Web-scraped datasets

The most prominent technique for collecting large in-the-wild facial datasets available to the research community is called *web scraping*. Web scraping refers to the automatized collection of data (in this case, face images and related metadata) from web sources. We continue by describing the most relevant examples of datasets constructed with this data collection technique.

The Images of Groups Dataset (GROUPS) [129] is a web-scraped collection of images containing groups of people, curated for the study of social context and contextual features. The web-scraping procedure consisted of querying the Flickr† image hosting search engine with search terms that have a high probability of yielding images of social groups (e.g., weddings and family portraits). Age and gender information were unknown, so the authors manually estimated age and gender categories. Age was categorized into seven groups: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, and 66+. The age distribution is not balanced, with the most represented

*https://images.google.com/
†https://www.flickr.com/search/

Table 4.1: Frequently used publicly available facial image datasets with age labels.

	General					Labels		Collection		
Dataset	Year	Images	Subjects	Im/Sub	Demographics	Type	Range	Images	Labels	Environment
FG-NET [35]	2002	1,002	82	12.22	unbalanced	chronological age	0-69	manual	manual	semi-controlled
CLAP 2015 [29]	2015	4,699	-	-	semi-balanced	apparent age	3-85	manual	manual	uncontrolled
CLAP 2016 [134]	2016	7,591	-	-	semi-balanced	apparent age	0-95	manual	manual	uncontrolled
Appa-Real [30]	2017	7,591	≈7,000	≈1.08	semi-balanced	chronological + apparent age	0-95	manual	manual	uncontrolled
MORPH-II [95]	2006	55,134	13,618	4.05	unbalanced	chronological age	16-77	manual	manual	controlled
AgeDB [135]	2017	16,488	568	29,03	semi-balanced	chronological age	1-101	manual	manual	uncontrolled
GROUPS [129]	2009	5,080	28,231	1.00	unbalanced	7 apparent age group	0-66+	web scraping	manual	semi-controlled
Adience [130]	2014	26,580	2,284	11.64	semi-balanced	8 apparent age groups	0-60+	web scraping	manual	uncontrolled
MegaAge [136]	2017	41,941	-	-	unbalanced	apparent age	0-70	web scraping	manual	uncontrolled
UTKFace [137]	2017	23,708	-	-	semi-balanced	apparent age	0-116	web scraping	DEX + web scraping	uncontrolled
AFAD [138]	2016	164,432	-	-	unbalanced	chronological age	15-40	web scraping	web scraping	uncontrolled
CACD [139]	2014	163,446	2,000	81.72	unbalanced	chronological age	14-62	web scraping	web scraping	uncontrolled
WIKI [89]	2016	62,328	62,328	1.00	unbalanced	chronological age	0-100+	web scraping	web scraping	uncontrolled
IMDB [89]	2016	460,723	20,284	22.71	unbalanced	chronological age	0-100+	web scraping	web scraping	uncontrolled
IMDB-WIKI [89][1]	2016	523,051	82,612	6.33	unbalanced	chronological age	0-100+	web scraping	web scraping	uncontrolled

[1] Separate entries for the IMDB and WIKI subsets were added to highlight the differences in their properties.

age groups being 20-36, while the gender distribution is fairly balanced. The collected dataset consists of 5,080 group images containing 28,231 faces. Images are collected in completely unconstrained conditions, offering a variety of head poses, facial expressions, occlusions, and lighting conditions. One of the big downsides is that the median face is very small, with only 18.5 pixels between the eye centers. Moreover, as the age group is manually estimated based on facial appearance, it represents the apparent age group rather than chronological age.

The Adience Dataset [130] was web-scraped to create a challenging in-the-wild age and gender estimation benchmark. In this case, Flickr was searched for mobile device photos with the Creative Commons (CC) license. Although the authors emphasize that this is an unconstrained collection of images, they discarded images of faces with estimated yaw head pose angles greater than 45°, as well as many challenging faces undetected by the simple Viola-Jones face detector [100]. Gender, apparent age group, and identity were manually labeled based on the images and the available contextual metadata related to them (e.g., image tags, captions, and image album information). Age was categorized into eight apparent age groups: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, and 60+. The dataset contains 26,580 images of 2,284 subjects. This is one of the few datasets with good distribution across age groups, including young subjects.

The Cross-Age Celebrity Dataset (CACD) [139] was the first public large-scale web-scraped dataset with chronological age annotations, initially introduced for cross-age face recognition. The goal was to create a large-scale dataset with good sample variety concerning age. The list of target subjects was created based on two main criteria: (1) the subjects should have varying ages, and (2) they must have large numbers of images available on the web. According to that, the authors decided to collect images of celebrities born in a 40-year period. They used the popular online movie database IMDb[‡] to find the 50 most popular celebrities born in each year from 1951 to 1990, totaling 2,000 subjects. The Google Image Search engine was used to scrape images. Search phrases combined celebrity names and specific years to collect samples with varying ages. The years from the search phrases were used in combination with the date of birth (DoB) information collected from the IMDb to automatically produce the age labels. This resulted in more than 160,000 samples. Although the authors admit that this simple approach produces a lot of noisy labels, the collected dataset was far superior to the existing ones in terms of size and sample variety.

The Asian Face Age Dataset (AFAD) [138] is another web-scraped dataset of similar size to CACD, with 164,432 images scraped from the RenRen Social Network (RSN), widely used by Asian students and graduates. The authors collected date of birth information provided by the users and images from special albums containing so-called *selfie* pictures, often containing only the user's face. The difference between the date of upload and the DoB was used as

[‡]www.imdb.com

chronological age ground truth. This approach relies on the assumption that the date of upload matches the date the photo was taken, that the users provided correct DoB information, and that selfie albums contain only single-subject photos of users. Being aware of this, the authors employed workers to manually filter out some obvious mistakes. Since most users were students, the age range is limited to 15 to 40. Moreover, the dataset contains only subjects from Asia, making it very biased and unsuitable for real-world generalization. The dataset also provides gender labels, with a female-to-male ratio of roughly 2:3.

The UTKFace Dataset [137] was initially collected by selecting images from the MORPH and CACD datasets and by filling in the distribution gaps for very young and elderly subjects with web-scraped images from Bing and Google search engines. The labels were obtained either from image captions or based on automatic age estimation of the DEX model [140]. This unconventional approach combines automatic apparent age estimation, web scraping, and manual involvement. Contrary to the 10,670 samples with uniform age and gender distribution mentioned in the reference paper, the publicly available version contains 23,708 samples and mentions only the web as the source of images. This in-the-wild dataset also provides labels for two gender and five ethnicity groups. The usability of this dataset heavily depends on the quality of pseudo labels generated by the automatic age estimation model.

The IMDB-WIKI Dataset [89] was collected with a similar approach to CACD, based on scraping images of famous people. The authors managed to collect more than 500,000 images with age and gender labels from IMDb and Wikipedia[§], making it the single largest public dataset for age and gender estimation to date. The authors used the IMDb to obtain a list of the 100,000 most famous actors and crawled images directly from their IMDb profiles, along with gender and DoB information. Additionally, they collected Wikipedia profile pictures with the same meta-data. After removing all the images that did not list the year they were taken, they used the listed years and the DoB from subjects' profiles to automatically obtain age labels. In the case of images with multiple face detections, they decided to keep only the images where all secondary face detection confidences were under a certain threshold. Similar to CACD, the authors note that they cannot vouch for the accuracy of the assigned age and gender information.

The MegaAge Dataset [136] is another example of a dataset consisting of images scraped from the Flickr platform. The dataset is made of samples randomly selected from the large in-the-wild MegaFace face recognition dataset [141] and the YFCC100M dataset [142]. Both MegaFace and YFCC100M are based on images scraped from the Flickr platform. The dataset consists of 41,941 images with age posterior distribution labels. The labels were obtained by manually comparing images to multiple annotated images from FG-NET. Annotators were asked to estimate if the person in the image appeared younger or older than the person in the reference image. The posterior reflects the apparent age with an estimated uncertainty range.

[§]https://en.wikipedia.org/

4.1.2 Video-based datasets

Similar to how essential FG-NET and MORPH datasets were for the progress of image-based age estimation research, publicly available video data is required to support the development of video-specific methods. Unfortunately, video datasets with age labels are far more scarce.

The UvA-NEMO Smile Dataset was initially introduced to study differences between spontaneous and posed smiles in [143]. Dibeklioğlu et al. used it for age estimation in [144] and defined an evaluation protocol. The dataset consists of 597 spontaneous and 643 posed smile recordings, totaling 1,240 videos. The videos were collected from 400 volunteers (185 female and 215 male) aged 8 to 76 years. Subjects were primarily Caucasian. The recordings were done in a controlled environment, with constrained illumination and high-resolution cameras. They used funny video clips to elicit spontaneous smiles, while instructional videos were used for posed smiles. Trained annotators selected and segmented a balanced number of genuine and posed smiles. Each segment starts with a neutral expression and transitions to a smiling expression.

The UvA-NEMO Disgust Dataset [145] was collected concurrently with the UvA-NEMO Smile Dataset, following the same recording setup. 324 volunteers (152 female, 172 male) were recorded posing disgust facial expressions. 313 of them also participated in the UvA-NEMO Smile Dataset collection. Similar to the Smile version of the dataset, age varies from 8 to 76 years, and the subjects are primarily Caucasian. Trained annotators selected 518 disgust video segments, where every segment once again starts with a neutral expression and transitions to a disgusted expression.

The UvAge Dataset [146] was created specifically for age estimation from videos. It consists of 6,898 videos from 516 subjects. They proposed a web-scraping technique to obtain videos with precise chronological age annotations. The approach was based on a collection of videos of celebrities with birth information available on Wikipedia. To get a reliable video recording time, they used traceable public events such as the Academy Awards or the G20 Summit. The videos were manually verified and segmented into sequences containing only one subject. The proposed celebrity web-scraped data collection approach resulted in a larger and more difficult in-the-wild dataset compared to the UVA-NEMO datasets. Along with age labels, each video was also annotated with identity, gender, ethnicity, and occupation.

The Casual Conversations Dataset (CC) [147] is a recently published video dataset from Facebook AI (now Meta AI) designed for measuring fairness of computer vision and audio models across a diverse set of ages, genders, apparent skin tones and ambient lighting conditions. It consists of 45,186 high-quality videos collected from 3,011 subjects. According to the authors, a distinguishing feature of this video dataset is the precise chronological age and gender annotations provided by the subjects themselves. Additionally, they labeled low ambient light videos while apparent skin tone is annotated by a group of trained annotators according to the

Table 4.2: Overview of the video-based age estimation datasets with chronological age labels.

Dataset	Subjects	Videos	Age range	Demographics	Head pose	Illumination
UvA-NEMO Smile [143]	400	1,240	8 - 76	unbalanced	mostly frontal	constrained
UvA-NEMO Disgust [145]	324	518	8 - 76	unbalanced	mostly frontal	constrained
UvAge [146]	516	6,898	16 - 83	unconstrained	unconstrained	unconstrained
Casual Conversations [147]	3,011	45,186	18 - 85	balanced	unconstrained	unconstrained
Casual Conversations Mini [147]	3,011	6,022	18 - 85	balanced	unconstrained	balanced
Casual Conversations v2 [149]	5,567	26,467	18 - 81	balanced	unconstrained	unconstrained

Fitzpatrick scale [148]. Another distinguishable quality of this dataset is its permissiveness; the dataset is collected from consenting subjects and is publicly available for evaluation purposes of both academic and commercial models. The authors also proposed a well-balanced subset of the CC dataset, denoted as Casual Conversations Mini (CCMini). The subset was formed by selecting one dark and one bright video per subject (when possible) to have a balanced lighting distribution, with a total of 6,022 videos.

The Casual Conversations v2 Dataset (CCv2) [149] was collected to improve the robustness and fairness of audio, computer vision, and speech models. It is a follow-up to the original CC dataset, curated with a special focus on geographical diversity. 5,567 subjects from 7 countries were paid to participate in data collection, whereas all participants in the original CC dataset originated from the same country (i.e., USA). The 26,467 collected videos amount to 320 hours of scripted and 354 hours of non-scripted conversations. Some of the labels, such as age, gender, language/dialect, disability, physical adornments/attributes, and geo-location, were self-provided by the participants. Trained annotators were used to additionally label for Fitzpatrick Skin Type [148], Monk Skin Tone [150], voice timbre, recording setup, and per-second activity. The dataset is designed primarily for evaluation purposes but can also be used for model training. However, it is noted that certain labels, such as age, gender, disability, and physical adornments/attributes, can not be used for training.

A summary of the reviewed datasets is presented in Table 4.2. Our efforts to receive access to UvA-NEMO Smile, UvA-NEMO Disgust, and UvAge datasets were unsuccessful, making the Casual Conversation datasets the only publicly accessible resource for video-based age estimation research. However, note that the CC datasets do not permit the training of age models.

4.2 Age estimation algorithms

Formal definitions of the most relevant estimation algorithms are given in Section 2.3. Following those definitions and the taxonomy presented in Figure 4.1, we continue by reviewing relevant work according to the age estimation algorithm type.

4.2.1 Regression

Regression is the most overt age estimation algorithm. It treats both age labels and predictions as continuous values and aims to learn the mapping directly. Due to its simplicity, it is one of the two most widely utilized algorithms, along with multi-class classification. However, it implies the goal of estimating age with the precision of a single year, making it a very difficult task.

Work presented by Lanitis et al. [35] is often cited as the first age estimation algorithm based on regression. The authors represented the aging pattern by a quadratic function and proposed multiple approaches for determining suitable aging functions for estimating age from unseen images. The best-performing approach was named the Weighted Person Specific (WPS) approach. However, this approach relied on various external information, such as gender, health, and living style. A purely appearance-based approach, named the Weighted Appearance Specific (WAS) approach, used the Mahalanobis distances between training and test images as weights and calculated the weighted sum of known aging functions to estimate age from unseen faces.

Another early work by Zhou et al. [151] presented a general algorithm for image-based regression. Its effectiveness was demonstrated on three image-based tasks, including age estimation. They considered three data-driven regression methods: Nonparametric Kernel Regression (NPR) [152], Kernel Ridge Regression (KRR) [152], and Support Vector Regression (SVR) [152]. The age-related evaluation was performed on the FG-NET dataset. The SVR method performed the best of the three baseline methods, while the proposed Image-based Regression (IBR) method outperformed all of them.

Yan et al. [153] formulated a regression problem utilizing Semi-definite Programming (SDP) and compared the performance of the proposed approach to Quadratic Regression (QR) [154] and Multilayer Percepton (MLP) [65]. According to experiments on FG-NET, the SDP approach outperforms QR and MLP. However, this approach is computationally expensive and not suitable for large datasets.

Suo et al. [155] proposed an age estimation framework that utilizes a hierarchical face model and specially designed sparse features. Their framework was evaluated in combination with four different age regression methods: Age-group-specific Linear Regression (ALR), Multilayer Perceptron (MLP), Support Vector Regression (SVR), and Logistic Regression (LR).

The experiments carried out on the FG-NET dataset showed that MLP outperforms the other considered methods.

Considering work based on classic machine learning and handcrafted features, many other noteworthy regression-based approaches were proposed. Guo et al. [156] proposed the Locally Adjusted Robust Regressor (LARR) and compared its performance to SVM and SVR. Fu et al. [157] and Fu and Huang [158] used manifold learning and framework based on the Multiple Linear Regression problem [159]. Guo and Mu [160] presented the Kernel Partial Least Squares (KPLS) regression model, Chao et al. [161] introduced age-oriented local regression algorithm that combines K Nearest Neighbors and SVR (KNN-SVR), while Cai et al. [162] proposed model based on Gaussian Process Regression (GPR). An abundance of proposed age regression algorithms is evident. The Support Vector Regression algorithm seems to be the most popular choice, as it was considered in many relevant works, such as [151, 155, 156, 161, 163, 164, 165].

Even though a multitude of age regression algorithms was presented over the years, most of the early results were overshadowed by the superior performance of methods based on deep learning. The well-established deep learning regression method leverages Mean Square Error Loss and provides much better results, as reported in [166, 167, 168]. A less frequently used option is to rely on Mean Absolute Error Loss, yielding comparable performance in [168]. A combination of deep models and classic regression algorithms is explored in [169] and [170]. The former used a CNN combined with an SVR regressor, while the latter used deep features and fusion of Random Forest (RF) [171] and SVR outputs to estimate age.

4.2.2 Classification

Compared to regression, the classification approach offers more flexibility with respect to the expected level of precision. The age classification problem can be formulated as a binary problem where, for example, the model only needs to distinguish minors from adults. On the other hand, the precision of a single year can be facilitated by covering some predefined age range with an appropriate number of classes (e.g., 101 classes for an age range from 0 to 100). We refer to the latter formulation as *exact age estimation*, while formulations with sub-year precision can be referred to as *age group classification*. Section 3.2 explains this categorization in more detail.

As exact age estimation is a challenging task even for contemporary methods, early research was focused on age group classification with a very limited number of classes. In a very early work by Kwon and Lobo [172], published almost three decades ago, age classification was formulated as the task of distinguishing three age groups: babies, young adults, and senior adults. In later work by Dehshibi et al. [173], age was categorized into four groups. More recent work on age group classification is built around the GROUPS and Adience datasets. Gallagher and Chen [129] proposed classification to 7 age groups in the GROUPS dataset, while Eidinger

et al. [130] defined 8 age groups for the Adience dataset.

The multi-class classification formulation can be applied to many different tasks from various research fields, resulting in diverse classification algorithms. An early comparison of classifiers for automatic age estimation was carried out by Anitis et al. [154]. They considered the Quadratic Model, Shortest Distance Classifier, MLP, and the Kohonen Self-Organising Map. The Quadratic and MLP approaches were performing best on a small private dataset. Hajizadeh and Ebrahimnezhad [174] applied the Probabilistic Neural Network (PNN) to the age classification problem, Wang et al. [175] utilized the Furthest Nearest-Neighbor (FNN) algorithm, Han et al. [176] used a multi-class AdaBoost algorithm [128], while Sawant et al. [177] considered Gaussian Process Classifier (GPC). Similar to how we identified SVR as the most frequently utilized approach in the case of classic regression methods based on handcrafted features, Support Vector Machine (SVM) seems to be the predominant choice for age classification [140, 165, 178, 179, 180].

Although CNN models were combined with SVM in [181] and [182], the deep learning method based on Cross Entropy Loss and Softmax seems to be the contemporary method of choice for age classification [137, 183, 184, 185]. Deep learning also enabled the use of some more complex methods, which are reviewed in the following section.

4.2.3 Advanced estimation algorithms

Based on various works reviewed in this section, neither regression nor classification is the optimal choice for the exact age estimation task. Regression is often claimed to be sensitive to outliers and errors in the data, while standard multi-class classification does not take into account the ordinal nature of age labels. Classification error will be the same regardless of whether the predicted class is adjacent to the ground truth class or if the prediction is very distant from it. For example, given that the ground truth is 20 years, predicting 19 and 80 results in equal classification error. The advanced methods reviewed in this work aim to solve these issues by combining regression and classification in various ways, by improving label representation and loss function formulation, or by designing more complex estimation frameworks.

Label distribution learning

Label distribution learning (LDL) methods aim to improve alignment with the ordinal nature of age labels by modifying label representations and loss function formulations. In the case of standard age classification, labels are formulated as basic one-hot vectors. For regression, the age label is simply a continuous value. The basis for label distribution learning algorithms lies in modifying label representation in a way that each training sample is not associated with a single label but with a distribution of labels. Each value in such distribution represents the degree to

which the corresponding label describes the training sample. This value should be highest for the ground truth label, while it gradually decreases as we move away from the ground truth. The difference between such modified label representation and conventional label representations is shown in Figure 4.3. This approach is also referred to as *soft classification*.

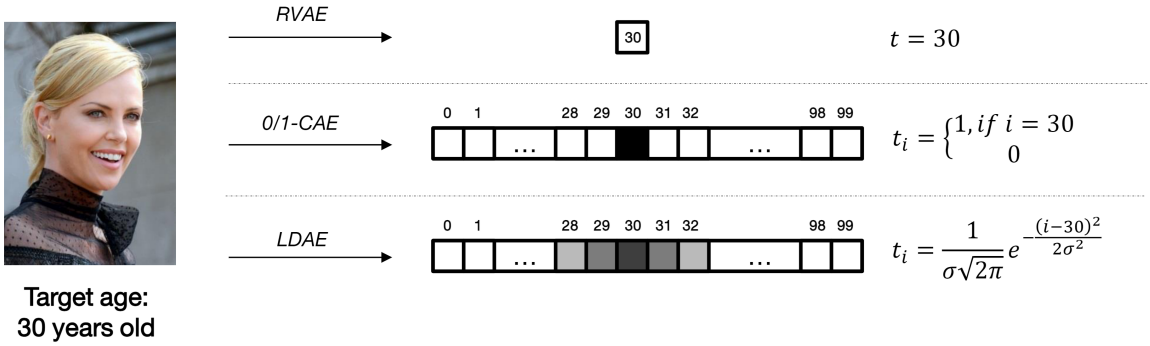

Figure 4.3: Label encoding for regression (RVAE), classification (0/1-CAE), and label distribution (LDAE), where t denotes the resulting encoding and σ is a LDAE hyper-parameter. Reprinted from [186] with permission from Elsevier.

Geng et al. [36] introduced the LDL concept to the age estimation field and considered the Triangle and Gaussian distributions to formulate labels. Their method was named IIS-LLD, where IIS stands for Improved Iterative Scaling algorithm [187]. According to them, switching from one-hot vectors to label distributions also improves training on smaller datasets, especially in case of gaps in ground truth age label distribution. Since the label distribution covers the ground truth class and its adjacent classes, one face sample also contributes to the training of its neighboring classes. They carried out experiments on the FG-NET dataset and compared the single-label, Triangle distribution, and Gaussian distribution approaches, concluding that the Gaussian LDL performs best. Their results on FG-NET were outperformed by the PLS-LLD method proposed by Zeng et al. [188]. The IIS algorithm was replaced by partial least squares regression (PLS) to perform multivariate multiple regression analysis.

Yang et al. [189] were the first to combine deep learning with label distribution learning and proposed the Deep Label Distribution Learning method. Their method shares some similarities with [36]. Namely, they utilized Gaussian label distributions and used Kullback-Leibler Divergence to formulate the loss function. However, their method relies on a multi-stream CNN-based architecture, and they modified the parameters of Gaussian distribution for each sample in the training set according to the standard deviation of apparent age votes. Their method was ranked 4th on the ChaLearn LAP 2015 apparent age challenge [29].

Work from [189] was extended for the ChaLearn LAP 2016 apparent age challenge [134] and dubbed Deep Age Distribution Learning (DADL). The LDL method mainly remained unchanged, while most changes were related to the deep learning model architecture and the ensembling approach. The DADL method was ranked 2nd on the challenge. Furthermore, Gao et al. [190] additionally extended the Deep Label Distribution Learning approach and dubbed

this method DLDL. They evaluated DLDL on age estimation, head pose estimation, multi-label classification, and semantic segmentation tasks, empirically demonstrating the robustness of the method. The label distribution method, however, remained mostly the same.

An age-difference-based approach was presented by Hu et al. [115]. Performance of a CNN-based age estimator trained on standard age datasets was improved by additional training on web-collected data with age difference information. The age difference was labeled as the difference between years when images were taken. The initial estimator was trained using the LDL approach. The loss was formulated as a combination of the entropy loss, cross-entropy loss, and K-L divergence distance to force the probability distribution of age predictions to have one single peak around the correct age.

To alleviate inconsistency between the training objectives and evaluation metrics in the DLDL framework, Gao et al. [91] extended it with an expectation regression module. The LDL and expectation regression modules are combined with a light-weight CNN and trained in an end-to-end manner, forming the DLDL-v2 method depicted in Figure 4.4. The proposed method was shown to outperform metric regression, ranking, deep expectation, and standard DLDL algorithms.

Later, Li et al. [191] argued that fixed-form age distribution is not suitable to represent complicated facial image domains. To mitigate this problem, He et al. [192] constructed label distributions by learning the cross-age correlation between context-neighboring samples, while Li et al. [191] proposed a label distribution refinery that adaptively learned the continuous age distribution.

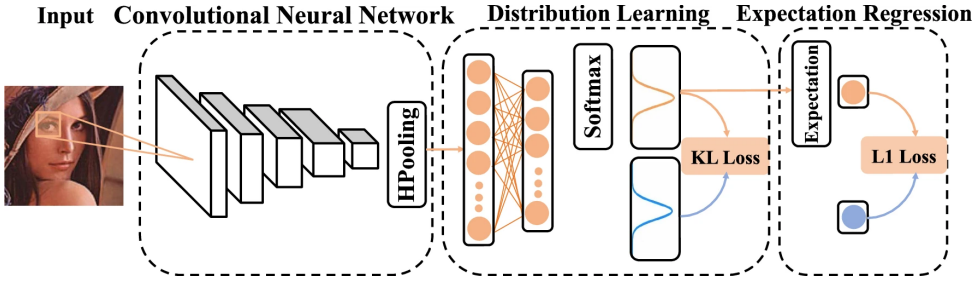

Figure 4.4: The framework of the DLDL-v2 method, combining label distribution learning and expectation regression. Reprinted from [193] with permission from Springer Nature.

Ordinal ranking

Another category of algorithms that aim to improve alignment with the ordinal nature of age labels is called ordinal ranking. The main idea of such algorithms is to divide a complex task into a series of simpler sub-tasks while taking into account the relative order of the labels. Predominantly, this is done using a series of binary classifiers that simply answer whether the age associated with the input x_i is higher than some threshold age k or not. In practice, this is

accomplished by another special type of label encoding and aggregation of K binary classifiers, assuming age range $[0, K]$. In such encoding, the first k values in the encoding vector are set to 1, while the rest is set to 0. This is visualized in the Figure 4.5, part (b). The age can then be calculated according to Equation 2.11.

The first to introduce such an algorithm to the age estimation field were Chang et al. [194, 195], building on the reduction framework from [196]. They reduced the inference problem to a set of simple cost-sensitive binary sub-problems and utilized the ordinal hyperplane ranking (OHRank) algorithm to solve it. They demonstrated the superiority of the proposed ranking method over multiple regression and classification methods on FG-NET and MORPH-II benchmarks. In their later work [197], they carried out a theoretical analysis covering the cost of each individual binary classifier and adopted a translation-invariant and deformation-stable scattering transform to extract facial features. Li et al. [198] adopted the OHRank algorithm and proposed a semi-supervised version in order to incorporate information embedded in unlabeled images and to further improve the OHRank performance. The used low-dimensional aging representation was learned to maximally preserve ordinal information of facial images and the underlying local structure information.

Yang et al. [199] were the first to combine the ordinal ranking approach with deep learning. They proposed combining extraction of facial features through a 3-layer scattering network (ScatNet), dimensionality reduction with PCA, and age prediction via a category-wise ranker. The encoding for each ranker was formulated as a combination of the category labels and the ordinal age rank. This allows the category-wise rankers to consider joined high-level representations while learning age prediction. They coined this approach as DeepRank+.

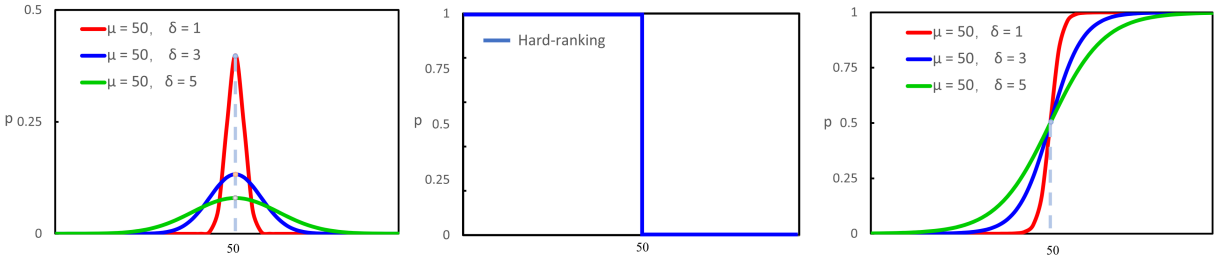

Figure 4.5: Label encoding in case of label distribution learning (a), hard ranking (b), and soft ranking (c) for value $k = 50$. Reprinted from [200] under CC BY 4.0 DEED license.

Later, Niu et al. [138] proposed OR-CNN, an end-to-end deep ranking method that jointly optimizes the feature selection and age prediction tasks. They used a multiple-output CNN that collectively solves a series of binary classification sub-problems. Their experiments demonstrated superior performance over CNN-based regression and OHRank methods on MORPH-II and the newly introduced AFAD datasets. Contrary to the single multi-output CNN approach from [138], Chen et al. [201] used a series of basic CNNs, each serving as a binary classifier for

a specific age rank. Their experiments demonstrated superior performance over the OR-CNN method on the MORPH-II dataset.

To combine the benefits of ordinal information and correlation between adjacent ages (leveraged by the LDL algorithms), Zeng et al. [200] proposed the Soft-ranking method. To achieve their goal, they introduced a novel age encoding method, depicted in Figure 4.5, part (c). They performed experiments on MORPH-II, AgeDB, CLAP 2015, and CLAP 2016 datasets, consistently outperforming other ranking methods.

Hybrid methods

There is a plethora of research that either combines some of the previously reviewed estimation algorithms or proposes fundamentally different algorithms that don't fall under any of the reviewed categories. We continue by briefly reviewing some of the hybrid algorithms that made a significant impact in the field.

There are a lot of instances where regression and classification were directly combined. Han et al. [34] evaluated the informativeness of different facial components and compared their results with human performance. Their hierarchical approach used a binary decision tree based on SVM to perform age group classification. Within each age group, a separate age regressor was trained to predict the final age. The regressors were trained with overlapping age ranges to mitigate the age estimation errors due to incorrect age group classification. Li et al. proposed AgeNet [202], a method that averages outputs of regression and classification subsystems. For this purpose, they used two separate CNNs. One network was trained with a combination of label distribution and modified Softmax Loss, while the regressor CNN utilized a Sigmoid layer and Euclidean Loss. An interesting way of combining distribution learning and classification systems was proposed by Antipov et al. [110]. To deal with low performance on ages under 12 years, a separate specialized CNN-based children age classifier was trained, while the main CNN used LDL-based training. If the output of the main CNN was below 12, they employed the CNN specialized for children. Otherwise, they used the expected value calculated from probabilities predicted by the main CNN. The use of expected values was introduced by Rothe et al. [89] in their Deep Expectation (DEX) method. The DEX networks were trained for classification with 101 output neurons for ages from 0 to 100, but the final continuous age predictions were regressed as a weighted average of the 101 Softmax outputs. According to experiments on the CLAP 2015 benchmark, this hybrid method outperformed both classification and regression counterparts.

As both ranking and label distribution learning methods were shown to be very successful, a lot of advanced variations that blur the lines between those algorithms were also proposed. A method proposed by Pan et al. [90] was based on a newly introduced loss formulation, coined Mean-Variance (MV). Although this approach does not rely on LDL-based label encoding, they

combined regression-based and distribution-oriented components in the loss function. They compared the performance of the MV loss to the classification-based Softmax and regression-based Euclidean loss on the MORPH-II dataset and concluded that a combination of Softmax and MV loss is performing best. A method that blurs lines between ordinal ranking and label distribution learning was proposed by Diaz and Marathe [203]. Their method is based on label encoding named Soft Ordinal vectors (SORD). The proposed method was tested on image quality ranking, age estimation, horizon line regression, and monocular depth estimation tasks, showing good generalization capabilities. Li et al. [204] proposed the use of an adaptive LDL approach (ALDL). Whereas the label distribution is predetermined and fixed in the standard LDL approach, their goal was to learn the distribution with an instance-aware adaptive technique. They argued that the MV loss, while also falling under the ALDL category, results in undesirable multimodal distributions due to the joint use of the Softmax and MV losses without required constraints. They proposed a novel Unimodal-Concentrated loss formulation that was shown to outperform Mean-Variance on the MORPH-II benchmark.

4.3 Image-based age estimation methods

To further review age estimation methods, we focus on image-based methods and review them according to the feature representation approach. As proposed by the taxonomy in Figure 4.1, there are two fundamentally different approaches to feature extraction. Early efforts in the field were oriented toward the manual design of novel discriminant features. These *handcrafted* feature extraction algorithms relied on domain expertise and careful engineering to transform the raw data to a suitable representation [63]. Contrary to the handcrafted features based on manual design and human ingenuity, feature extractors based on learning are optimized to extract discriminative representations directly from raw data without explicit instructions and engineering. Deep learning has become the dominant paradigm in recent years due to the ability to learn multi-level representations and very complex functions from millions of samples, thus achieving superior generalization performance.

4.3.1 Handcrafted feature representation

Conventionally, five distinctive approaches to feature representation modeling are placed under the umbrella of handcrafted features. Those are anthropometric representations, active appearance models, aging pattern subspace, age manifold, and texture representations. The parameters of these feature extractors are usually manually tweaked to extract as much age-related information as possible, including head shape, wrinkles, and skin texture. Generally, handcrafted models are less complex than their deep-learning counterparts and thus require less data and

computational power to achieve good results. However, models based on handcrafted features are usually restricted to frontal head poses, can not efficiently leverage large amounts of images, and are generally less accurate.

Anthropometric representations

Anthropometric measurements are used to quantify the overall structure of the human body and dimensions of its features such as bones and muscles [42]. The study of measuring proportions and sizes of human faces is called face anthropometry. Farkas [205] relied on 57 facial landmark points to define five measurements: shortest distance, axial distance, tangential distance, angle between locations, and angle of inclination, resulting in a total of 132 facial measurements. To a limited extent, face anthropometry measurements can help us distinguish different age groups.

Kwon and Lobo [131, 172], often cited as authors of the first work on automatic age estimation, used anthropometric measurements and density of wrinkles to distinguish babies, young adults, and senior adults. Six distance ratios between facial landmarks were used to separate babies from adults, while snakelets were used to extract wrinkle patterns from skin areas in order to separate young adults from senior adults. Similarly, Horng et al. [206] used a combination of geometric and wrinkle features to classify subjects into four age groups, including babies, young adults, middle-aged adults, and old adults. They used two neural networks, where the first one used geometric features to determine if the subject was a baby or not, while the second one used wrinkle features obtained by Sobel filtering to classify the subject into one of the three adult groups. Dehshibi and Bastanfard [173] further relied on a combination of distance ratios between landmarks and a back propagation neural network to classify subjects into four age groups. Ramanathan and Chellappa [207] proposed a craniofacial growth model based on eight distance ratios defined over facial landmark points to model age progression in minors under 18. Turaga et al. [208] further explored the role of geometry in age estimation. The space of facial landmarks was interpreted as a Grassmann manifold and the exact age estimation problem was posed as a problem of function estimation on the manifold. The same geometric features were used by Thukral et al. [209], where a hierarchical approach was used to first divide subjects into various age groups and then regress the exact age based on group-specific regressors. Al-Shannaq et al. [3] note that anthropometric features are not appropriate for non-frontal facial images and that relying solely on geometry dismisses important appearance features (e.g., skin texture) that are more appropriate in the case of adult subjects.

Active appearance models

Active appearance models (AAM) are statistical models introduced by Cootes et al. [210]. This image representation, commonly used for facial images, combines anthropometric and texture-based model descriptors. The shape and texture features are extracted by a dimensionality

reduction algorithm based on a set of images. Commonly, PCA is used to produce a parametric face model. Lanitis et al. [35, 154] introduced AAM to the age estimation field. They first extended AMM to the age estimation problem and observed a correlation between 50 raw face model parameters and the actual age in [35], then evaluated different types of classifiers based on AAM features in [154]. Later, Yan et al. [211] used AAM to explore a ranking approach with uncertain labels, Chen et al. [212] used AAM in combination with a novel cumulative attribute concept for regression learning in case of sparse and imbalanced data, while Feng et al. [213] used it to explore age estimation by cost-sensitive label ranking and trace norm regularization.

According to Angulu et al. [4], AAM has a clear advantage over anthropometric models that consider shape features only, as additional texture features make AAM appropriate for age modeling at all stages of life. However, ElKarazle et al. [42] note that dimensionality reduction required for the AAM approach can still lead to important aging features, such as wrinkles, going unnoticed.

Aging pattern subspace

Geng et al. [23, 214] based their work on AAM and introduced the Aging Pattern Subspace (AGES) method. The AGES method defines the aging pattern as a sequence of sorted face images belonging to a single subject. As such aging patterns are often incomplete, AGES compensates for missing elements in the sequence by learning a subspace representation of the subject's images. The age is estimated by positioning an unseen image at various locations in the pattern and selecting the aging subspace that has minimal image reconstruction error. PCA is used to obtain the subspace representations, and reconstruction error is minimized by iterative expectation maximization (EM). Geng et al. [23, 214] evaluated this approach on the task of exact age estimation on the FG-NET benchmark and reported superior performance. However, Angulu et al. [4] note that the applicability of AGES is hindered by the need for datasets with face images at several different ages for every subject. Moreover, as AGES relies on AAM, it shares the same drawbacks in terms of not utilizing some important aging features such as wrinkles.

Age manifold

Whereas AGES finds a specific pattern for each person, the age manifold method, introduced by Fu et al. [157], finds a trend for multiple subjects at different ages. Moreover, this method is more flexible than AGES as it allows for a subject representation to be based either on a single image or multiple images at different ages, thus simplifying the data collection problem. The age manifold method learns the common low-dimensional pattern from multiple faces at every age. Fu and Huang [158] formulated a manifold using Conformal Embedding Analysis (CEA), while age was estimated using multiple linear regression. Scherbaum et al. [215] applied

manifold learning to a 3D morphable model, where a manifold was formed by isosurfaces of nonlinear SVR function. Guo et al. [163] replaced PCA dimensionality reduction with Orthogonal Locality Preserving Projections (OLPP), which is a supervised manifold learning algorithm. Yan et al. [216] proposed Synchronized Submanifold Embedding (SSE), where manifold learning is dually supervised by subject identity and 3D head pose. Experiments on FG-NET indicated superiority over conventional regression and unsupervised manifold learning algorithms. More recently, Cai et al. [162] proposed the Discriminative Gaussian Process Latent Variable Model (DGPLVM), an effective probabilistic model for manifold learning, where the mapping between low-dimensional representations and ages was found by Gaussian process regression.

Texture representations

The texture-based models are appearance models that directly depend on the pixel values of facial images. These models rely on various texture operators to extract skin features related to aging, such as spots, lines, or edges. Local Binary Patterns (LBP) [119] are one of the most effective and frequently used texture descriptors, utilized for age estimation by [217, 218, 219], among many. Here, every pixel in the image is simply represented by a binary code formed by comparing the intensity of a central pixel to its surrounding pixels, making it computationally efficient. Gao and Ai [220] extracted Gabor features of 3 scales and 4 orientations. The resulting 12 magnitude images were used as raw features, while fuzzy LDA was used to classify age. They reported performance superior to raw pixels and LBP. Guo et al. [165] proposed an age estimation method based on Biologically Inspired Features (BIF) [221], a popular type of feature inspired by the primate visual cortex. They used a variation of BIF particularly designed for age estimation, utilizing Gabor filters to model receptive fields and MAX and STD operations as sources of nonlinearity. El Dib et al. [222] extended the BIF approach from [165] by incorporating fine detailed facial features, automatic initialization using Active Shape Models (ASM), and analyzing the complete facial area, including the forehead details. Instead of extracting BIF from a holistic face like in [165], Han et al. [34] also extracted BIF features from individual facial components, such as forehead, eyebrows, eyes, nose, and mouth. An important advantage of BIF features is that they can effectively handle small translations, rotations, and scale changes [3]. In addition to these popular options, several other handcrafted texture features were successfully applied to the age estimation problem. A feature descriptor called Spatially Flexible Patch (SFP) was utilized by Hayashi et al. [223] and Belver et al. [224]. Zhou et al. [151] used Haar-like feature extractors, Histograms of Oriented Gradients (HOG) features were used by Hajizadeh and Ebrahimnezhad [174], while scattering transform descriptors were evaluated by Chang and Chen [197]. According to ElKarazle et al. [42], the texture-based models are more apt to perform well on images taken in uncontrolled conditions.

4.3.2 Learned feature representation

While there are many ways in which feature representations can be learned, the prevailing approach used in the age estimation field is to learn representations directly from the raw texture (i.e., image pixel intensities). This is similar to the use of handcrafted texture representations described in the previous sections. However, learning feature representations is a fundamentally different approach, as the representation is discovered directly from the data. The Multilayer Perceptron (MLP) was one of the first methods used to directly learn from the raw pixel data. The potential of this approach was not fully discovered during the early adoption years and was widely abandoned. However, another category of neural network architectures was later introduced specifically to address the problem of learning feature representation directly from images: Convolutional Neural Networks (CNN). The introduction of CNNs to the age estimation field resulted in consistent improvements across the field, making CNN architectures the dominant design choice in the deep learning era. Huerta et al. [166] were the first to apply deep learning to age estimation and compare the results to the traditional approaches under the same experimental settings. In an exhaustive set of experiments, HOG, LBP, and SURF handcrafted feature extractors were compared to a simple LeNet [73] CNN architecture variant, demonstrating the clear superiority of the CNN-based model. According to Al-Shannaq et al. [3], the performance differences exhibited by the CNN-based models are related to two factors: the choice of CNN architecture and the choice of pretraining procedure. While CNN models can be initialized with random weights and trained directly on the target data (i.e., *from scratch*), the full potential of deep learning models is unlocked by the transfer learning technique, briefly introduced in Section 2.2.3. In line with this, we continue reviewing deep-learning-based age estimation methods according to the choice of CNN architecture and the type of pretraining procedure.

Neural network architectures

CNN is the most popular deep learning architecture, widely used across different computer vision fields. Various advanced CNN architectures have been proposed over the last decade, and many of them have been successfully adopted in the age estimation field. Bianco et al. [225] performed an in-depth benchmark analysis of 20 prominent deep neural network architecture families. For a detailed overview of the CNN architecture choices, we refer readers to this study while we continue by briefly discussing the most popular choices in the age estimation field.

The earliest example of a CNN feature extractor used for age estimation is LeNet [73]. This very simple architecture, based on only two convolutional and two fully connected layers, was used in early work by Huerta et al. [166], and later by Dong et al. [226]. The AlexNet [77] architecture gained its reputation as the first winner of the ImageNet challenge [86]. This

slightly deeper network, consisting of 5 convolutional layers and 3 fully connected layers, was successfully utilized for age estimation in [181, 227, 228, 229]. The next network architecture to gain popularity was VGGNet [78], along with its many variants: VGG-11, VGG-13, VGG-16, and VGG-19. These architectures were designed to be much deeper, with the premise that more nonlinearities can be modeled with an increased number of layers. The most popular version, VGG-16, has 13 convolutional and 3 fully connected layers. Along with VGG-Face, a specialized version pretrained on almost 1,000,000 face images for person identification, it is the most widely used architecture family in the age estimation field [20]. It was used for age estimation by [32, 90, 91, 110, 189, 203, 204, 230], among many. The same year as VGGNet was introduced, Google introduced GoogLeNet [79]. This is an even deeper architecture, with up to 22 convolutional layers, but with much fewer parameters due to the removal of heavy, fully connected layers. Based on its newly introduced inception module, this network is also referred to as Inception-v1, while several additional Inception versions were introduced in subsequent years. GoogLeNet was utilized for age estimation in [170, 202, 231]. While the trend of stacking more and more convolutional layers was obvious, a limitation related to the vanishing gradient problem emerged, requiring a novel approach. ResNet [80] architectures proposed an answer to this obstacle, based on residual blocks with skip connections that enable gradient propagation in very deep networks. While this enabled training of networks with more than 200 layers, the ResNet-50 variant turned out to be the most popular one. ResNets were used for age estimation in [232, 233, 234]. MobileNet [83] is a more recent family of architectures oriented towards achieving comparable performance under limited computational resources. The main difference, compared to the ResNet architecture, was the introduction of efficient depthwise separable convolutions. MobileNets were used for age in [235, 236]. In addition to these popular CNN architectures, some less popular but specialized networks were also used for age estimation, including ScatNet [237] in [199], C3AE in [238], SSR-Net in [236], and ThinAgeNet and TinyAgeNet in [91].

While the CNN architecture is undoubtedly the dominant choice in the field, it is also important to note that other types of neural networks are being successfully used for age estimation. For example, Zhang et al. [239] combined traditional CNN backbones with the Long Short-Term Memory (LSTM) network architecture to extract local features of age-sensitive regions. Recently, the Visual Transformer (ViT) architecture, originating and widely used in the natural language processing (NLP) field, showed excellent results compared to the well-established CNN architecture. Paplhám and Franc [240] compared the FaRL backbone [241], based on a ViT-B-16 model [242], to the popular VGG-16, ResNet-50, and EfficientNet-B4 CNNs under a unified framework, concluding that no backbone emerges as universally best across all age estimation datasets.

Pretraining procedure

While the deep networks described in the previous section enable impressive results, they also require large training data. Deep networks have great capacity and can model very complex nonlinear functions. However, in the case of a small and unbalanced training dataset, this can easily lead to overfitting issues, meaning that the learned nonlinear function closely fits the training set but does not generalize well to unseen samples. In case of a sufficient amount of task-specific data, the models can be trained from scratch without the use of external data and pretraining steps. If the target task-specific dataset is not sufficiently large, pretraining and transfer learning techniques can mitigate the overfitting issues and improve the generalization performance.

Due to the above issues, a relatively few researchers trained their networks from scratch. LeNet is a very simple architecture, utilizing only two convolutional and two fully connected layers. Huerta et al. [166] were able to train it from scratch based on relatively small datasets. Yan et al. [243] designed a 5-layer network and trained it from scratch for several tasks, including age estimation. However, they did not manage to outperform a method based on handcrafted BIF features. To prevent overfitting issues, Levi and Hassner [185] also proposed a shallow CNN architecture, utilizing only 3 convolutional layers. Moreover, they inflated the training set by applying cropping augmentations. Wan et al. [244] also relied on shallow networks with 3 convolutional layers, organized in cascades to use auxiliary demographic information in age estimation, and trained them from scratch. We can see that in the case of training done from scratch and based on relatively small datasets, only shallow networks seem to be successfully utilized.

Three common pretraining techniques were proposed to fully leverage the potential of deep networks by using large amounts of external data, as reviewed in [20]. The first technique is called general task pretraining (GT), where the model is pretrained on some general task unrelated to age estimation or even to faces. The most popular option for GT pretraining is the ImageNet [86] dataset. A typical ImageNet pretraining includes classification training with 1,000 classes and more than a million training images. Due to the nature of deep learning, the initial hidden layers learn generic low-level features, such as points, edges, and contours. The hidden layers of GT pretrained networks are not optimized for age estimation but are general enough to be useful in the age-specific finetuning step. The ImageNet GT pretraining was used for age estimation in many publications, such as [186, 189, 231, 245, 246]. The second technique is called face recognition pretraining (FR). Instead of learning to classify general objects, the FR pretrained models are trained to recognize identity from facial images, thus learning face-specific representations. The most popular dataset for FR pretraining is the VGGFace2 dataset [247], consisting of more than 3 million images. Another popular option is the CASIA-WebFace dataset [248], consisting of about 500,000 facial images. The FR pretraining was used

for age estimation in [170, 181, 183, 202, 230, 245, 249]. The third technique is called age estimation pretraining (AE). Here, models are pretrained on large facial datasets designed for age estimation. This option best suits the task at hand as it enables models to learn very specialized features. However, it requires a large dataset with age labels, which is difficult to acquire. The IMDB-WIKI dataset, introduced specifically for the purpose of AE pretraining by Rothe et al. [89], consists of more than 500,000 web-scraped weakly labeled noisy samples. Even though the quality of this dataset is very poor, it enabled considerable performance improvements in [32, 110, 140, 244, 250]. Antipov et al. [186] performed a comparison of GT and FR pretraining techniques and concluded that FR is more suited for the age estimation task. Carletti et al. [20] further commented that AE pretraining shows the greatest potential, especially if better-suited AE pretraining datasets become available.

4.4 Video-based age estimation methods

One of the main premises of this thesis is that facial videos provide extensive information that is useful for age estimation. However, compared to the fruitful image-based research field, there is only a handful of published work that leverages video information to improve age estimation. As described in an early work focused on analyzing facial behavioral features from videos by Hadid et al. [251], there are two main strategies for video-based face analysis. The simplest strategy is to apply image-based methods to all video frames or a set of sampled frames from a video. Individual frame results are then fused over the sequence. The more elaborate strategy involves leveraging both face appearance and face dynamics information through spatiotemporal modeling. We review video-based methods with a focus on spatiotemporal modeling and with respect to the type of used feature representations.

4.4.1 Handcrafted feature representation

To compare the aforementioned strategies, Hadid et al. [251] implemented two baseline methods. The image-based baseline used Local Binary Pattern (LBP) features [252, 253] extracted from static images, followed by SVM classification. A similar video-based method combined Volume-LBP features [254] selected using AdaBoost and SVM classification, with the goal of using both static facial information and facial dynamics. They trained and tested their age classifiers on a set of 2,000 web-scraped videos that were manually annotated with apparent age labels. Age estimation was treated as a classification task with 5 age classes. Although their experiments indicated that the video-based approach can improve performance for face recognition, gender classification, and ethnicity classification tasks, their image-based method outperformed the video-based counterpart in the case of age estimation.

A similar line of exploration was done by Dibeklioğlu et al. [144], with a focus on the discriminative power of smile dynamics for age estimation. They leveraged movement features of facial key points, such as speed, acceleration, and amplitude, to complement appearance-based cues and improve estimation accuracy. Facial dynamic features were combined with LBP features extracted from a single frame and fed into SVM classifiers and regressors. Their experiments were based on the UvA-NEMO Smile Database [143], which contains 1,240 videos from 400 subjects. Each video shows a single subject's transition from a neutral expression to a posed or spontaneous smile. Their extensive experimentation demonstrated significant performance improvement over the baseline image-based method, as well as compared to the video-based method from Hadid et al. [251].

Dibeklioğlu et al. extended their work on face dynamics for age estimation in [145]. In addition to LBP appearance features, they considered Intensity-based Encoded Ageing Features (IEF) [255], Gradient-based Encoded Ageing Features (GEF) [255], and Biologically-inspired Ageing Features (BIF) [165]. Surface area features based on a mesh model were used instead of facial key-point movement features. Appearance and dynamic features were selected and combined in different ways, followed by a newly introduced two-level classifier, where the age ranges for classifiers were adaptively selected. Along with the previously used UvA-NEMO Smile dataset, they introduce the UvA-NEMO Disgust dataset. Their experiment showed that video-based methods that leverage face dynamic features significantly outperform the image-based baseline on both Smile and Disgust versions of the UvA-NEMO datasets.

4.4.2 Learned feature representation

Instead of handcrafted features used in [144, 145, 251], a combination of CNN, RNN, and attention modules was used by Pei et al. [256] to improve performance of age estimation from facial expression videos. The proposed Spatially-Indexed Attention Model (SIAM), depicted in Figure 4.6, used a CNN for appearance modeling, a spatial attention module for the detection of salient facial regions, an RNN model for capturing facial dynamics, and a temporal attention module for temporal saliency. The model is trained in an end-to-end manner. To train and validate their method, UvA-NEMO Smile and UvA-NEMO Disgust datasets were once again used, following the protocol from [145]. Their experiments demonstrated that the proposed neural-network-based approach outperforms previous methods based on handcrafted features on the UvA-NEMO Smile dataset. However, those improvements were not successfully replicated on the smaller UvA-NEMO Disgust dataset. The authors attribute this to a relatively small size of the UvA-NEMO Disgust datasets, arguing that neural-network-based methods require larger amounts of data in general.

Ji et al. [257] pointed out that deploying image-based age estimation models directly to videos often suffers from estimation stability issues. To address this, they proposed a combina-

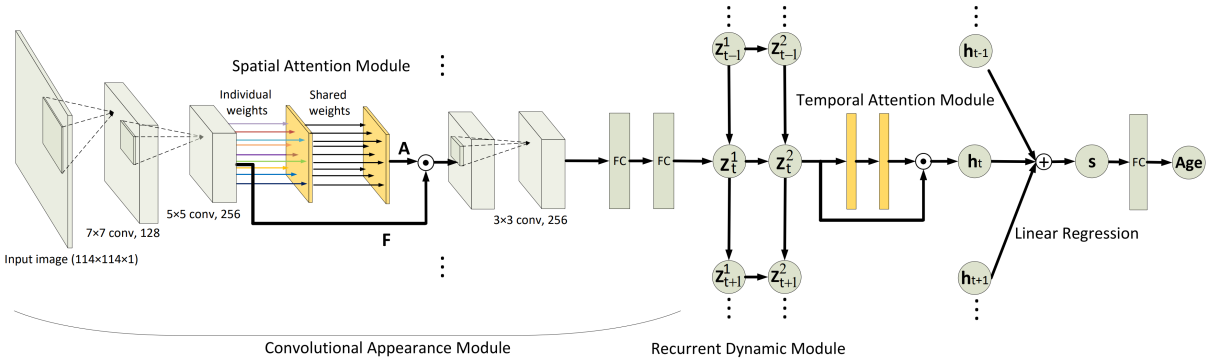

Figure 4.6: Spatially-Indexed Attention Model (SIAM) for video-based age estimation. Reprinted from [256] with permission from IEEE (© 2020 IEEE).

tion of CNN-based feature extraction and attention-based feature aggregation modules, sharing some similarities with [256]. Their model was not trained in an end-to-end manner, as the CNN was initially trained on images, followed by attention module training. Along with the commonly used MSE loss, their loss function for the attention module included a component for estimation stability based on mean estimation calculated on the sequence level. The feature extraction module was trained on the MORPH-II [95] image dataset. To train the feature aggregation module, they built a video dataset comprising 18,282 frames from a single twelve-minute facial video of one subject, noting that there are no available age facial video datasets. Their experimental results demonstrated improvements concerning both age estimation accuracy and stability.

To attenuate the effect of head-pose on video-based age estimation, Han et al. [146] based their method on pose-invariant *uv* texture maps reconstructed from video frames by a Wasserstein-based GAN. They introduced the UvAge dataset consisting of 6,898 videos from 516 subjects and used it both for training and evaluation. The proposed AgeGAN method, which simultaneously inpaints the partial *uv* maps and estimates age from them, was shown to perform better than methods based on original 2D face images and methods based on different types of inpainting modules. Even though video data was used in their experiments, the proposed method was focused on achieving robustness to head pose changes without exploiting any temporal information from videos.

Zhang et al. [258] also focused their work on mitigating the negative effect of head-pose variation on age estimation from videos. The proposed method relies on multiloss CNN for head-pose estimation [259] and Deep Regression Forests (DRF) [260] for age estimation. They trained the head-pose estimation model on the 300W-LP dataset [261] and used it to create multiple subsets of the CACD [262] and AFAD [138] datasets based on different angle thresholds. The selected subsets were subsequently used to train DRFs for age estimation. After selecting the best threshold, CACD and AFAD datasets were divided into frontal and non-frontal subsets. The performance of models trained and tested on frontal subsets was shown to be significantly

better. They proceeded by collecting two 12-minute facial videos from two subjects for evaluation purposes, also noting that no video datasets are available for age estimation model training. The head-pose estimation was used to cherry-pick frames with near-frontal faces from video sequences, while non-frontal faces were ignored. Their experiments showed that this frame cherry-picking step improves both age estimation accuracy and stability. Similar to the work from Han et al. [146], their method was focused on mitigating head-pose-related performance drop without exploiting any temporal information from videos.

4.5 Discussion

We conclude this section with a holistic overview of the current state of the age estimation field. As this is a very fruitful research field, a plethora of datasets, algorithms, and methods were proposed, many of which were reviewed in this section. The taxonomy of age estimation system design choices, presented in Figure 4.1, helps us grasp the diversity of the proposed solutions. When performing such a review, a common goal is to try to determine the strengths and weaknesses of different algorithms and methods and, finally, the best-performing option for the task at hand. Typically, we should also aim to discover potential gaps in the field and opportunities for novel contributions. We continue by summarizing the benefits and drawbacks of the reviewed approaches, discussing the options for the *state-of-the-art* comparison, and commenting on what motivated us to pursue our contributions presented in Chapters 5 and 6.

As reviewed by Carletti et al. [20], it is a consolidated conviction that deep learning offers overwhelming superiority over dated methods based on handcrafted features and classic learning algorithms. When considering the use of video inputs instead of images, we can conclude that the benefits of leveraging spatiotemporal dynamic information are evident. However, one needs to take into account that video-based age estimation is an understudied field with a certain lack of available public data. The question of the best estimation algorithm type is the most difficult one to answer, as there is a large variety of choices to consider, each having some distinctive downsides. Classification and regression are indeed the simplest and most popular choices. However, regression formulates the aging process as a linearly growing dependence while also being sensitive to outliers and data errors. Classification, on the other hand, does not take into account the ordinal nature of age labels and is prone to overfitting issues in the case of gaps in data distributions. The basic ranking algorithm was reported to be suboptimal due to inconsistency in the training objective and the evaluation metric. While this is not definitive, and there are many advanced methods that blur the line between these algorithm categories, the label distribution learning methods are often reported to have the most competitive performance [2, 110, 186, 240]. The strengths and weaknesses of different algorithm types are reviewed in more detail in [28].

Conventionally, researchers pursue a state-of-the-art (SOTA) comparison to determine what is the best performing method for a certain task. This is usually done by compiling results reported in various publications while trying to put them under a common denominator, thus facilitating a fair comparison. Extensive SOTA reviews are compiled in [2, 3, 4, 20, 43, 263]. A prerequisite for a fair comparison is a well-defined and publicly-available benchmark. As discussed in Section 2.4, a benchmark protocol consists of benchmark data, performance metrics, and the evaluation protocol determining how to calculate the defined metrics on the given data. The protocol needs to be defined in a way that makes training and testing data fully independent. Moreover, the test data should ideally be a well-balanced representative of the target real-world environment. Problematically, this is not the case for the most frequently used benchmarks: FG-NET and MORPH-II. The demographic distribution of the MORPH-II dataset is presented in Figure 2.6).

Over the years, we have put a lot of effort into the reproduction of various methods and, more often than not, faced reproduction issues. Many of the issues originated from ill-defined benchmark protocols. Important protocol details, especially regarding data splits, are often left out. Furthermore, as stated at the beginning of this chapter, details of the complete age estimation framework, including face detection, preprocessing, feature extraction, and attribute prediction, are in many cases not fully disclosed. While analyzing various reproduction issues, we formed the opinion that details regarding the used data and the early data processing steps play a role that is too significant to be overlooked. Interestingly, this is fully aligned with the conclusions of the recent work by Agbo et al. [28]. As details of these steps are frequently left unspecified, we question the feasibility of carrying out a genuinely fair SOTA comparison. This opinion was further corroborated by inconclusive method comparisons, reviewed in [20]. More significantly, Zeng et al. [200] empirically proved that the problem of overlapping identities between the training and testing data in popular age benchmarks gives rise to misleading results.

A very recent study from 2024 by Paplhám and Franc [240], accepted for publication in the most prestigious conference in the field (Conference on Computer Vision and Pattern Recognition), focuses solely on a call to reflect on the evaluation practices for age estimation. In this unconventional work, a comparative analysis of SOTA methods was performed under a unified framework designed to facilitate fair comparison. They isolated the influence of the tested method by keeping other relevant components in the system, such as data collection, data processing, model design, training, and evaluation, constant. They compared 9 estimation methods over 7 datasets in the cross-dataset evaluation setting. Unexpectedly, the findings showed that performance differences between various methods introduced over the last decade are almost negligible, even though gradual and consistent improvements were reported in the related literature. Their experiments showed that the improvements are, in fact, stemming from

other factors, such as the amount of data used for pretraining, image resolution, model architecture, face alignment approach, and the amount of facial context used in the pipeline. They identified two trivial yet persistent issues. First, the reviewed benchmark protocols do not use standardized and publicly available data splits in most cases. Even more troubling is the finding that only $\approx 10\%$ of reviewed papers used the subject-exclusive protocol on the MORPH-II benchmark, meaning that some subjects were used both in training and testing in the majority of publications. Secondly, a clear mapping between the reported performance gains and various modifications in the estimation system often does not exist, as multiple components are changed simultaneously. Due to this, the authors of this study deem regular SOTA comparisons meaningless.

Interestingly, even though this was not our main objective in Chapter 5, the evaluation setup and the captured results of our exploration across 5 methods and 10 dataset versions are aligned with findings from [240]. Moreover, Paplhám and Franc identified the amount of the pretraining data as the most influential factor on model performance. This supports our decision to focus our contributions to image-based age estimation (Chapter 5) on the design of a new, large training dataset, instead of pursuing method improvements directly. The review also states that the CLAP 2016 dataset, which is used for the recent Appa-Real benchmark, is the only public option with standardized data splits and reliable labels suitable for age estimation. This is aligned with our own decision to use it as our main image-based benchmark. On a similar note, our contributions to video-based age estimation (Chapter 6) are further focused on solving data issues. Specifically, we focused on establishing a well-defined, publicly available benchmark protocol for video-based age estimation, and mitigating the issue of video training data shortage via a semi-supervised learning approach.

Chapter 5

Unsupervised biometric data filtering for refined age estimation

Years of research in the field have led us to believe that data is the cornerstone for the design of robust age estimation models with good generalization capabilities. As biometric data collection becomes an increasingly sensitive issue, the research community struggles with the collection of large amounts of reliable data for biometric tasks such as gender, age, and ethnicity estimation. For over a decade, image-based face analysis research relied on small, manually collected datasets, ranging from 1,000 to 50,000 samples. Geng et al. [36] stated that the lack of sufficient training data is one of the main challenges of facial age estimation, also highlighting problems in the collection of data with a wide age range. Carletti et al. [20] pointed out that the currently available datasets are not yet ready for real-world exploitation. They state that the disparity of performance in the cross-dataset setting is symptomatic of the fact that the datasets are not representative of real-world complexity. They also state that the collection of actual age labels for real-world data is a difficult task. A very recent evaluation of SOTA methods under a unified framework, carried out by Paplhám and Franc [240], has shown that performance differences across various age estimation methods are negligible compared to performance gains obtained by the use of pretraining on large amounts of data. In fact, they concluded that pretraining data is the most influential factor in the design of an age estimation system.

Several research groups have recently utilized automatic web-scraping methods to successfully collect large amounts of noisy but heterogeneous data, and improve the state-of-the-art facial analysis algorithms [89, 115, 264, 265]. Today, the majority of researchers rely on large and noisy, web-scraped datasets [20]. Although a small amount of noise in the training data is not considered to be a problem for modern deep learning algorithms and can, in some cases, even help to reduce overfitting problems, large amounts of noise can reduce the smoothness of the cost function hyperplane, lower the convergence rate, and impair the final performance.

This chapter is based on our work from [266]* and [267]†, focusing on *picking out the bad apples* from facial image datasets. The goal of this work is to automatically reduce the level of label noise in web-scraped facial datasets by filtering out the wrongly labeled or otherwise faulty samples in order to improve the resulting face analysis algorithms. Again, we are focusing on the age estimation task, as it is one of the most difficult face analysis problems [134]. The contributions of the work presented in this chapter are summarized as follows:

- We present an efficient unsupervised method for biometric data filtering that can significantly reduce label noise in facial image datasets. It is an automatic and parameter-free method for facial dataset filtering that does not require supervised training of dataset-specific systems but utilizes only general-purpose, off-the-shelf algorithms and models.
- We apply the proposed filtering method to two state-of-the-art, web-scraped datasets, and demonstrate its benefits to 5 different age estimation methods, and with respect to generalization capabilities in unconstrained conditions.
- We propose a biometric filtering strategy to reinforce and refine the merging process of multiple facial datasets and derive the new Biometrically Filtered Famous Figure Dataset (B3FD). We demonstrate B3FD's superiority over existing state-of-the-art age estimation datasets with respect to both real and apparent age estimation and make the dataset publicly available.
- We highlight the importance of training data quality compared to the training data quantity and demonstrate that the proposed refinements of the training data result in a larger margin of improvement than the utilization of more advanced age estimation methods.

The rest of the chapter is organized as follows. Building on the related work review presented in Chapter 4, Section 5.1 reiterates relevant web-scraping techniques and reviews the most relevant dataset filtering methods. Further, Section 5.2 describes the proposed method for unsupervised biometric data filtering and provides experimental validation of its effectiveness. Section 5.3 describes the design strategy of the new famous figure age estimation dataset and provides a comparison with the state-of-the-art. Section 5.4 discusses the findings of this work.

5.1 Filtering web-scraped facial data

A review of work related to facial age estimation is presented in Chapter 4, with a special focus on web-scraped data in Section 4.1.1. To reiterate, web scraping refers to the automatized data collection from web sources. In this specific case, it refers to the collection of face images and related metadata. Section 4.1.1 introduced many relevant web-scraped datasets, such as GROUPS [129], Adience [130], CACD [139], AFAD [138], UTKFace [137], IMDB-WIKI

*Reproduced with permission from SciTePress.
†Reproduced with permission from Springer Nature.

[89], and MegaAge [136]. Their most important properties are summarized in Table 4.1. Design principles for all these datasets are fairly similar, with some nuances described as follows.

The main component of a facial age estimation dataset is the image data. For the aforementioned web-scraped datasets, images were collected via the Flickr platform in [129, 130, 136], Google Image Search in [137, 139], Bing search engine in [137], Internet Movie Database (IMDb) in [89], Wikipedia in [89], and RenRen Social Network in [138]. The second and equally important component of a dataset is the labels. Several different approaches for automatic label collection were proposed. When using image search engines, images are collected by querying the engine with search phrases. In [139], such search phrases combined celebrity names and specific years. The specified year was used in combination with the date of birth (DoB) information collected from IMDb to produce the age labels automatically. This technique relies on the assumption that queries will result only in images taken in the specified year. In the case of the AFAD dataset, collected from the RenRen Social Network in [138], self-provided DoB was used in combination with the date of image upload, expecting that this date also matches the date when the image was taken. In the case of the IMDB-WIKI dataset [89], the authors used a large list of famous actors from IMDb and scraped data directly from their IMDb profiles, including images, gender, and DoB information. They also used Wikipedia profile pictures with the same metadata. By using only images that listed the year in which they were taken, they were able to automatically obtain age labels by subtracting the DoB information from the listed year. All these techniques also rely on the assumption that the person of interest will be the only (or the main) subject in the scraped image.

Additional two examples of facial data web-scraping that did not result in a publicly available dataset followed similar basic principles. A very simple yet effective method for the automatic collection of a sizeable web-scraped gender dataset was presented in [264]. By querying search engines with a list of gender-specific names, the authors collected 4 million weakly labeled samples and demonstrated the importance of large-scale datasets for in-the-wild gender estimation. To avoid the need for large-scale public datasets with exact age annotations, a method for web-based collection of samples with age difference labels was proposed in [115]. To build their dataset, the authors used the Flickr platform and names from the LFW dataset [268] to crawl large amounts of images along with descriptions containing dates of image acquisition. Although they did not collect the actual age information, pretraining their network for age-difference estimation improved their final age estimation results.

Web-scraped datasets such as CACD and IMDB-WIKI were shown to be superior to the manually collected datasets in terms of size and sample variety, but their overall quality is undermined by the high amounts of label noise. We continue by reviewing efforts made toward cleaning noisy web-scraped facial datasets.

An early example of an automatic facial dataset filtering method was presented in [265].

In an attempt to design a robust and universal age estimator, the authors used image search engines and a set of age-related queries to collect a large facial dataset with weak age labels. To reduce the label noise levels, they designed a simple two-step filtering approach. In the first step, they used parallel face detection based on multiple state-of-the-art face detectors. To remove non-facial images and dismiss misaligned detections, they only retained samples with multiple detections overlapping more than 90%. To further reduce the number of false positive detections and to reduce the number of faces not correctly corresponding to the search-query age, they applied PCA to all images collected for a certain age and dismissed images with large reconstruction errors.

The age-specific PCA filtering step was intended to remove age-category outliers based on their apparent age, but the largest reconstruction errors were actually caused by face occlusions and non-frontal head poses, thus removing samples crucial for training a robust age estimator. Furthermore, due to the strict criterion for multiple face detection overlap, an additional large number of valuable difficult samples was discarded.

Even though the benefits of pre-training on the large and noisy IMDB-WIKI dataset were clearly demonstrated in [89], a cleaned version could further improve their age estimation results. In order to create a cleaned version of the dataset, a combination of automatic and manual processing steps was used in [110]. In the first step, all the images with multiple face detections were removed to increase the probability of the detected face corresponding to the provided age label. In the second step, a subset of the remaining multi-face images was manually filtered via a crowdsourcing annotation process.

The authors state that the first step ensures the correctness of the age labels. Still, both false positive and false negative detections induce considerable amounts of label noise, even in single-detection images. In the manual step, the annotators were asked to pair the provided annotation with one of the faces in the image. A study on human performance showed that the average annotator estimates age with a high mean absolute error of 4.7 - 7.2 years [34], indicating that even this seemingly trivial step can produce additional noisy outcomes.

Compared to a limited amount of work presented on age data filtering, several more advanced approaches for facial dataset filtering were proposed in the facial recognition field, as it has become one of the most data-hungry image analysis fields in general.

A data-driven approach for cleaning large face datasets was presented by authors of the FaceScrub dataset [269]. To identify the faces to be removed from their dataset, they exploited the observations that the same person should appear at most once per image, have the same gender, and look similar. The task of outlier detection was formulated as a query-specific quadratic programming (QP) problem based on a combination of terms related to those observations. Assuming that falsely detected faces form only a small portion of the detected set, they were able to train a one-class SVM and use its output as a score for the false positive term. To enforce

a gender term, they trained a two-class linear SVM for gender classification with query-based gender labels. Similar to the false detection term, the outputs of its decision function were used as gender scores. A similarity term was encouraged by graph regularization based on the normalized graph Laplacian, and an additional prior term was used to encode the assumption that most faces are correct.

By manually annotating a part of their dataset, the authors assessed their algorithm and demonstrated that their QP formulation outperforms the naive approach where the classifiers were used separately. However, the discussed benefit of manual workload reduction was somewhat impaired by the need for the dataset-specific classifier trainings.

The more recent large-scale web-scraped facial recognition dataset, named VGGFace2 [247], adopted and improved a multi-step semi-automatic approach from the original VGGFace paper [88]. To achieve their goal of a 96% pure dataset, their efforts included more than three months of manual annotations. The majority of that time was spent on the initial name list filtering. The annotation team reduced the initial list from 500,000 to only 9,244 names by dismissing all subjects for whom the top 100 Google Image Search results were not at least 90% pure. After applying a relatively strict face detection step, a set of 1-vs-rest classifiers was trained to discriminate between the 9,244 subjects. The threshold was selected by manually checking results for 500 subjects, and all samples with scores below the chosen threshold were dismissed. The next step, designed to remove near-duplicate images, used VLAD descriptor clustering and retained only one image per cluster. To detect overlapping subjects (names referring to the same person), an additional classifier was trained to generate a confusion matrix and remove classes mostly confused with others. The final, partially manual step consisted of iterative retraining of the 1-vs-rest classifiers with an annotator team manually filtering only part of the samples based on the classification scores.

To reach their target in terms of data purity, the authors of the VGGFace2 trained several versions of more than 9,000 1-vs-rest classifiers, trained an additional classifier for overlap detection, performed manual threshold search and substantial amounts of manual filtering. This impressive data filtering effort resulted in a state-of-the-art face recognition dataset.

5.2 Unsupervised biometric data filtering

To design an efficient filtering method, an overview of which is given in Figure 5.1, we analyzed the common sources of label noise in the current state-of-the-art biometric facial datasets.

Due to the nature of the commonly used web-scraping approaches described in Section 5.1, there are two main sources of label noise. The first problem is the unreliability of the automatic age annotation process itself. Although the date of birth (DoB) information is mostly correct, the year of image acquisition can be inaccurate or misleading. For example, Rothe et al. [89]

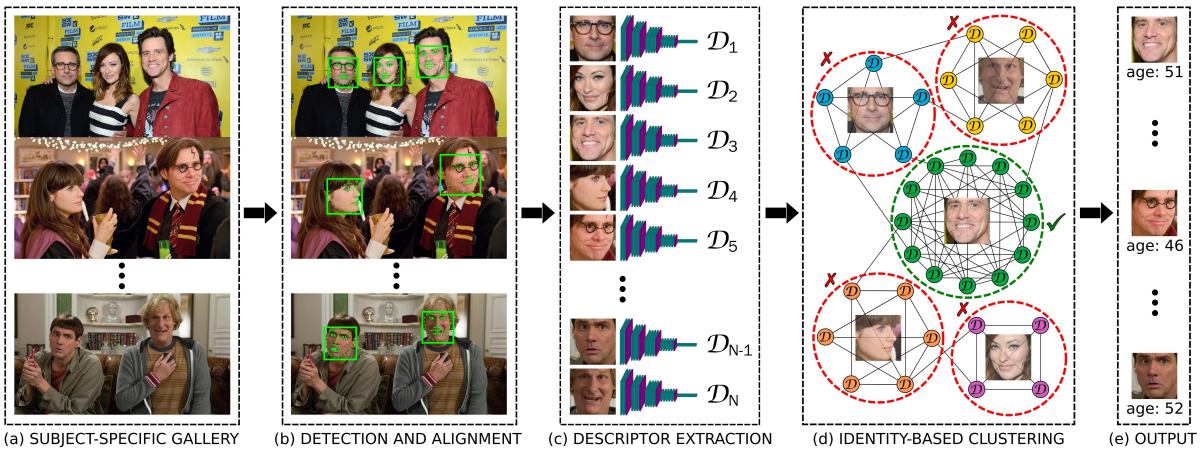

Figure 5.1: The proposed unsupervised biometric filtering pipeline for the IMDB subject Jim Carrey. (a) Samples are grouped into subject-specific galleries based on the provided identity metadata. Every image in the subject-specific gallery contains the gallery owner (e.g., Jim Carrey). Still, many also contain other subjects that can be mismatched with the gallery owner's biometric labels (e.g., age and gender). (b) Face detection and alignment of faces in the image gallery. Most of the detected faces belong to the gallery owner, while others are sources of label noise. (c) CNN-based face recognition descriptor extraction from all detected faces. (d) Graph-based clustering of the extracted facial descriptors. Clusters are formed based on identity matching. The gallery owner's cluster is the largest because their face appears most frequently in the image gallery. (e) Samples belonging to the largest cluster are retained, while those from the other clusters are discarded as noise.

mention that a large number of images are actually movie screenshots annotated with the year of the movie release, while some movies have production time spanning over several years. This problem usually causes only minor age annotation errors.

A much more serious issue, causing large discrepancies for age and other biometrics labels, is mismatched identities. In case of multi-person images, face detector failures, or bad image search results, the metadata can be paired with a face detection of a wrong subject. Let's consider an image containing a famous actress and her son, where the metadata is linked to the actress. If the son's face gets detected as the primary face, potentially due to the actress's face being partially occluded or recorded under a difficult angle, the sample will end up with a wrongly assigned gender label and a high age annotation error. (i.e., up to several decades).

Figure 5.2 shows examples of correctly paired and mismatched images for one subject appearing in both CACD and IMDB-WIKI datasets. Compared to typically constrained manually collected data, the top-row (i.e., correct) samples exhibit superior variation concerning many important aspects such as head pose, facial expression, lighting, and background. On the other hand, the bottom-row (i.e., mismatched) samples greatly impair the overall quality and usability of the dataset.

In order to reduce the number of labeled samples with erroneously matched facial images and to mitigate this most detrimental source of label noise, we propose a filtering method described in the next section.

Figure 5.2: Web-scraping noise in IMDB-WIKI and CACD data for subject Gal Gadot. For each dataset, the top row shows valid samples, while the bottom row shows image samples that have been wrongly paired with Gal Gadot's metadata.

5.2.1 Proposed filtering method

The main idea of the unsupervised filtering method is to automatically group samples from subject-specific image galleries (i.e., images with the same identity metadata) into clusters of samples with matching biometric descriptors and to keep only the samples from the largest cluster while all other samples get discarded. This way, the samples of the most frequently appearing subject in the image gallery can be retained, while samples erroneously matched with that subject's metadata are removed. This fundamental idea is expanded as follows.

A prevalent approach for automatically grouping samples based on a certain property is to utilize clustering algorithms. A number of clustering algorithms can perform the required grouping efficiently and without supervision, but regardless of the type of the clustering algorithm, the clustering performance will greatly depend on how the sample's grouping property is represented numerically.

For good performance, numerical representations should be compact and highly descriptive of the property of interest (e.g., the subject's identity). In the case of facial image data, a favorable option for the task is the facial recognition algorithms. Facial recognition algorithms are specifically designed to project high-dimensional facial image data to a highly discriminative low-dimensional biometric feature vector (i.e., face descriptor) that encodes the subject's identity. However, note that the proposed method is not restricted to descriptors obtained by facial

recognition algorithms, as other biometric or image descriptors can be utilized.

To reduce the undesired effects of feature extraction from misaligned and inconsistent detections, we propose to employ a two-step detection procedure consisting of regular object (face) detection followed by feature point detection that allows precise calculation of bounding box position and scale, as well as in-plane image alignment.

For the approach to be completely parameterless and unsupervised, the descriptor grouping should be done with a clustering method capable of automatically discovering the number of underlying groups (i.e., identities). For this purpose, a number of clustering algorithms, such as Chinese Whispers Clustering [58], Affinity Propagation Clustering [59] or Mean Shift Clustering [60], can be used.

The aforementioned outlines the basic concepts of the proposed unsupervised biometric filtering method, while the following section gives the implementation details of the designed filtering pipeline.

Implementation details

Based on the previously described concepts, we implemented a filtering system consisting of five main consecutive steps. A graphical summary of the proposed filtering pipeline is given in Figure 5.1, depicting the five steps for one specific subject. The implementation details of the presented steps are as follows.

Subject-specific gallery. Firstly, according to the provided identity metadata (e.g., subject name or ID), the datasets are reorganized into a series of subject-specific galleries. Each gallery contains all images collected for one specific subject. While the gallery owner should be present in all images in the gallery, many other subjects may appear as well, but less frequently.

Detection and alignment. The second step amounts to detecting and aligning all faces in the subject-specific image galleries. Although the information on the bounding box is usually provided, the given bounding boxes lack consistency concerning the bounding box scale and positioning. To ensure more consistent inputs to the following descriptor extraction step, we first utilize a face detection algorithm based on DLIB's[‡] CNN face detector to redetect faces and then use a facial alignment algorithm robust to bounding box imprecisions [270] to precisely determine bounding box position and scale based on the detected facial landmark points. Moreover, the facial landmark points are used to perform in-plane image alignment. The bounding box information provided by the dataset is used only in the rare cases of face detection failure, and even then, it is corrected by the face alignment step.

Descriptor extraction. The third step is the extraction of face descriptors for all the faces detected in the previous step. We utilized the DLIB's powerful facial recognition model based on the ResNet architecture [80] to extract compact 512-dimensional identity descriptors. By

[‡]http://dlib.net

calculating distances between the extracted face descriptors, the probability of two descriptors representing the same subject (i.e., identity) can be efficiently estimated, and by using DLIB's default descriptor similarity threshold, a reliable identity matching can be achieved. Well-performing face recognition systems offer reliable identity matching regardless of facial expression, orientation, and even occlusion to some extent, as well as environmental factors such as lighting and background.

Identity-based clustering. The fourth step aims to distinguish the samples of the gallery owner from other faces coincidentally present in the gallery, which may be incorrectly paired with the gallery owner's metadata. This step groups facial images based on their biometric features (i.e., identity) by applying a clustering algorithm to facial descriptors extracted from all the detected faces. Since the number of identities in the gallery is unknown, we utilize the Chinese Whispers clustering; an efficient graph-based parameter-free clustering algorithm introduced in [58], which discovers the number of clusters in a simple iterative process.

Output. For each subject-specific gallery, the pipeline is finalized by retaining facial samples and associated labels only from the largest identity cluster. All other samples, belonging to clusters associated with subjects appearing less frequently than the gallery owner, are discarded.

The designed filtering pipeline is completely parameterless and utilizes only generic off-the-shelf models and algorithms. This approach does not require supervised training of dataset-specific models or any type of manual effort.

5.2.2 Dataset filtering

The proposed filtering pipeline is applied to the two largest publicly available facial age estimation datasets: the CACD dataset and the IMDB-WIKI dataset. In this section, we discuss the method's prerequisites, data preprocessing, and the results of the dataset filtering based on the proposed method.

Prerequisites

There are two prerequisites that need to be satisfied for the proposed filtering method to be applicable:

1. There must be multiple images of every subject.
2. For each subject-specific image gallery, the number of appearances of the gallery owner's face must exceed the number of appearances of any other subject.

As we can see from Table 4.1, the average number of images per subject is 81.72 for the CACD and 22.71 for the IMDB dataset, indicating that the method's first prerequisite is satisfied for the majority of subjects. The WIKI subset of the IMDB-WIKI dataset has only one image per subject, so we will omit it from the filtering.

The probability of a well-defined image search producing more bad than good results is very low, especially for famous subjects. The probability of a subject not being the most frequently appearing person on its IMDb/Wikipedia profile photos is even lower. Therefore, the second prerequisite is satisfied intrinsically for the majority of samples from the CACD and IMDB-WIKI datasets.

Preprocessing

Although the training data can, in some cases, be used in its raw form for the supervised training of machine learning models, in most cases, a set of initial processing steps is applied to the data. Initial processing of the image data typically consists of image cropping, alignment, and filtering according to attainable image properties. Based on the label data, samples can also be filtered to remove invalid labels, improve sample distribution, etc. The goal of this section is to distinguish the three versions of data used in our experiments: raw, processed, and filtered.

Raw data (R). Both CACD and IMDB-WIKI datasets provide pre-cropped facial images with associated age labels. As face detection and cropping are the minimal preprocessing steps in typical face analysis systems, we consider this the raw data. This data consist of loose, unaligned face crops of varying quality and resolution, produced based on face detections obtained by the dataset's authors. The associated labels are most likely the direct product of the automatic labeling procedures, so invalid labels are likely to be present in the data. To make this raw data compatible with our training framework, we apply center-cropping to the non-square images, resize all images to the same resolution, and limit the age labels to the $[0, 100]$ range by applying function $age = min(max(age, 0), 100)$. None of the samples are discarded. This results in CACD-R, IMDB-R, and WIKI-R dataset variants.

Processed data (P). The processed data refers to the outcome of the typical automatic preprocessing steps, designed to improve the raw data quality and remove some of the obvious outliers that disrupt the learning process. Whereas the originally provided face detections are used in the raw data (R), the processed data (P) consist of re-detected, aligned, and re-cropped samples. The detection and alignment procedure is described in Section 5.2.1. Images that were damaged, had very low resolution, or in any other way caused the described detection, alignment, and cropping pipeline to fail were discarded. Additionally, a small number of samples that had age labels with biologically impossible (i.e., negative) or highly improbable (i.e., greater than 100) values were also discarded since they are most probably the result of failed automatic labeling procedures.

Filtered data (F). The filtered data are the data obtained by applying the proposed filtering method described in Section 5.2.1 to the previously described processed versions of the CACD and IMDB datasets (i.e., CACD-P and IMDB-P). This way, the filtered data (i.e., CACD-F and IMDB-F) is directly comparable to the processed data, and the impact of the proposed filtering

method is unambiguously observable.

Data augmentations, such as random cropping and image flipping, are not considered part of the initial processing as the same augmentation techniques are applied to all three versions of the data (i.e., R, P, and F) during the training procedure.

Filtering results

The raw versions of the CACD and IMDB datasets have 163,446 and 460,723 samples, respectively. By applying the initial processing, described in the previous section, we produced the CACD-P dataset with 150,383 samples and the IMDB-P dataset with 451,571 samples. After the proposed filtering method was applied to the processed data, 130,571 samples were retained from the CACD-P dataset (13.2% reduction), and only 216,595 samples from the IMDB-P dataset (52.0% reduction), giving us the final CACD-F and IMDB-F dataset versions. As we can see in Figure 5.3, the sample distributions of the filtered subsets of the CACD and IMDB datasets remained similar to the raw and processed versions, while the number of samples was greatly reduced.

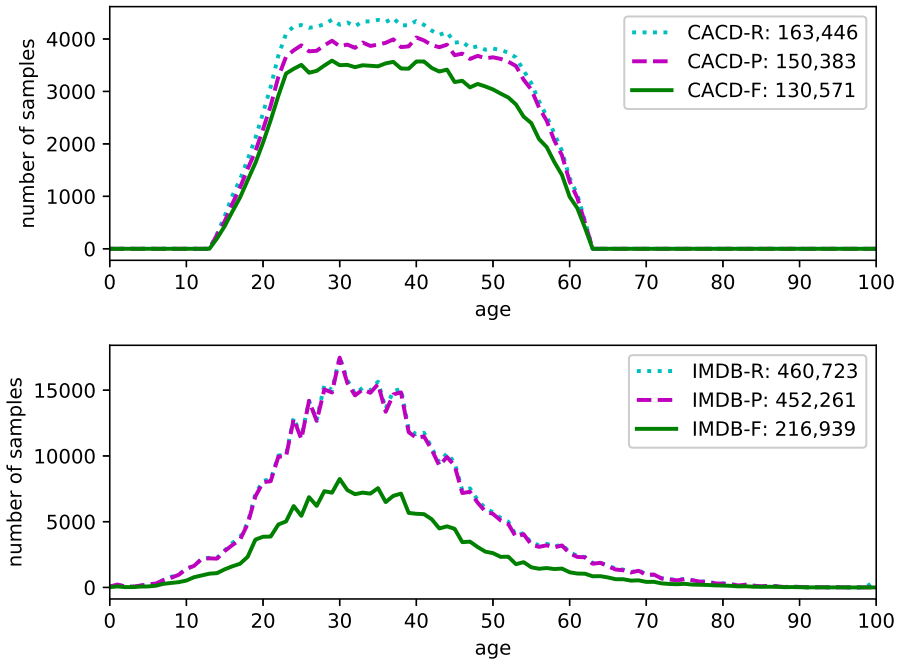

Figure 5.3: Age label distributions for the raw (R), processed (P), and filtered (F) versions of the CACD and IMDB datasets.

Several subject-specific galleries were manually inspected to examine the filtering results more closely and showed consistent results. Figure 5.4 shows the results of a statistical analysis of filtering outputs for one of the subjects from the IMDB dataset. The figure contains a histogram of the top five sample clusters and a chart representing the cluster sizes. The 48% of the samples that were grouped into the largest cluster were kept, while 52% of the samples were

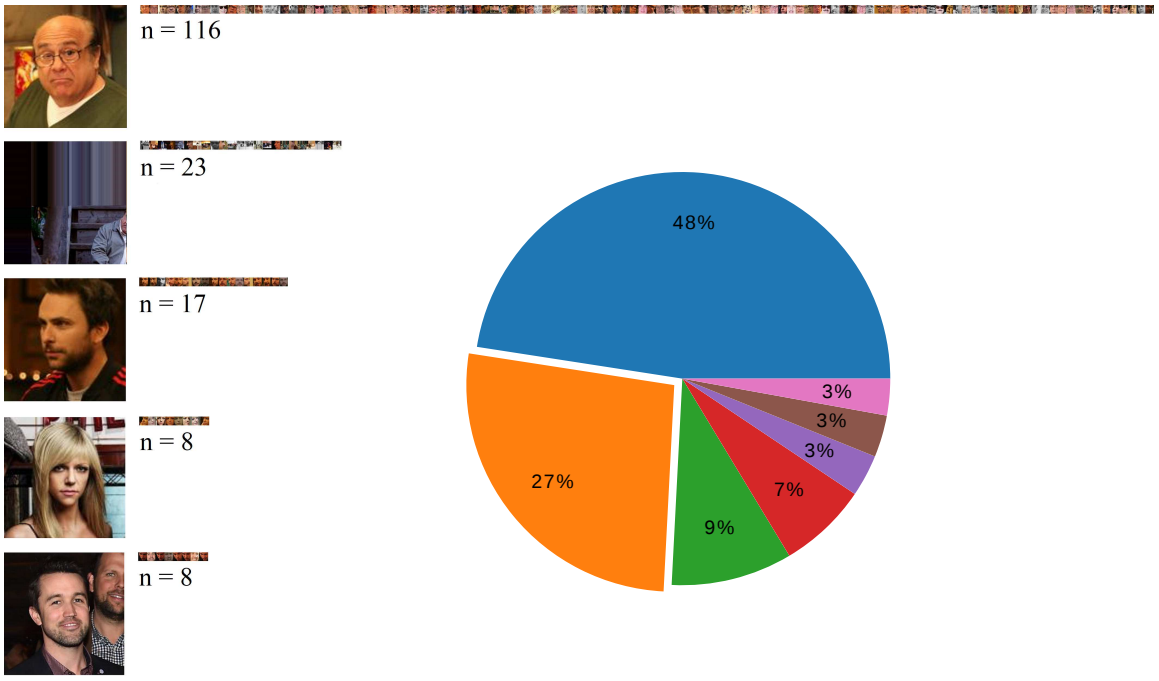

Figure 5.4: Clustering results for the IMDB subject Danny DeVito. Images on the left show representative samples for the top 5 clusters, horizontal bars contain all cluster samples (best viewed in electronic version with extreme zoom), and the chart on the right shows cluster size distribution with the emphasized part representing all clusters containing 1 to 3 samples [266].

filtered out. The analysis showed that the second largest cluster (9%) grouped primarily non-facial images caused by false-positive detections, and the subsequent clusters contained facial images of actors with whom the subject is most frequently associated. The emphasized part of the chart in Figure 5.4 represents all clusters with only 1 to 3 samples (1 to 3 occurrences per identity), grouping the less frequently appearing outliers.

5.2.3 Experimental evaluation

The proposed filtering pipeline described in Section 5.2.1 resulted in strong sample count reduction, as presented in Section 5.2.2. To validate that the resulting subsets of the original datasets have higher proportion of valid samples and that the proposed automatic filtering approach benefits the datasets' applicability to the facial biometric task they were designed for, we perform an extensive set of age estimation experiments.

Good generalization capabilities, crucial for real-world in-the-wild applications, often directly depend on the training data size and diversity. To validate that our aggressive sample reduction does not impair the generalization capabilities of the trained models, we performed cross-dataset testing on the unconstrained, manually collected Appa-Real benchmark. According to [20, 240], this is currently the most suitable benchmark for age estimation, as it offers reliable labels, heterogeneous images, and well-defined data splits. To validate if the benefits are method-invariant, we perform the evaluation based on two frequently used and three more

advanced age estimation methods, as described in the following section.

Age estimation methods

To perform unbiased evaluation of different age estimation methods, we used an identical CNN feature extraction model for all methods, thus supporting each of the methods with the same number of learnable parameters. The models were designed to output 101-dimensional feature vectors intended to facilitate the estimation of 101 age values ranging from 0 to 100. Based on this setup, we implemented five relevant age estimation methods, formally introduced in Section 2.3. Those methods are the MSE-based regression, the Softmax-based classification, the Deep Expectation (DEX) method from [89], the Mean-Variance (MV) method from [90], and the Deep Label Distribution Learning (DLDL-v2) method from [91].

In addition to the formal definitions of these methods from Section 2.3, it is important to specify certain hyperparameters related to two of those methods. In Equation 2.8, related to the Mean-Variance method, the factor λ_1 controls the contribution of the Mean Loss component, while λ_2 controls the contribution of the Variance Loss. Following the setup from [90], we set λ_1 and λ_2 to 0.2 and 0.05, respectively. The DLDL-v2 method combines the distribution-based Kullback-Leibler Divergence Loss and the regression-based L1 Loss. The parameter λ from Equation 2.9, which controls the balance between the DLDL-v2 loss components, was set to 1 in all experiments. The Equation 2.10 is used to calculate the distribution-based label encoding from the ground truth label. The σ hyperparameter, controlling the shape of the distribution, was set to 2.

Experimental setup

To compare results before and after the proposed filtering method was applied, we train the models on the processed (P) and the resulting filtered (F) versions of the datasets under identical conditions. Additionally, we train the baseline models on the raw data (R) under the same setup.

A set of image data augmentations, consisting of randomized horizontal flipping (i.e., image mirroring), bounding box perturbations, and in-plane rotations, was applied to training subsets of all three versions of the dataset. Random in-plane rotations were limited to $[-10^\circ, 10^\circ]$ range. The bounding box perturbations were achieved by resizing the image to 110×110 pixels and performing random cropping to 96×96, effectively obtaining perturbations of up to 15% of the bounding box scale.

The model selected for this extensive evaluation was a simple 9-layer CNN model based on the open-source architecture Tiny DarkNet[§]. This minimalistic 1M-parameter architecture, proposed for real-time performance in [271], was pretrained on the face recognition task and

[§]https://pjreddie.com/darknet/tiny-darknet/

Table 5.1: Results of the 5-fold validation of the raw (R), processed (P), and filtered (F) versions of the CACD and IMDB datasets. The testing and fine-tuning were performed on the Appa-Real benchmark. R-MAE and A-MAE denote mean absolute errors for real and apparent age estimation, respectively.

		Validation	Testing		Fine-tuning	
Dataset	Images	R-MAE	R-MAE	A-MAE	R-MAE	A-MAE
CACD-R	163,446	5.56 ± 0.03	14.60 ± 0.58	13.15 ± 0.67	9.11 ± 0.59	7.39 ± 0.72
CACD-P	150,383	4.66 ± 0.01	13.61 ± 0.13	12.30 ± 0.15	7.64 ± 0.17	5.78 ± 0.13
CACD-F	130,571	$\mathbf{3.46 \pm 0.01}$	$\mathbf{11.83 \pm 0.26}$	$\mathbf{10.60 \pm 0.26}$	$\mathbf{7.17 \pm 0.13}$	$\mathbf{5.46 \pm 0.15}$
IMDB-R	460,723	8.11 ± 0.03	11.14 ± 0.11	9.51 ± 0.08	9.84 ± 0.37	8.12 ± 0.37
IMDB-P	451,571	7.70 ± 0.02	8.51 ± 0.18	7.20 ± 0.14	6.73 ± 0.20	5.20 ± 0.34
IMDB-F	216,595	$\mathbf{5.07 \pm 0.01}$	$\mathbf{6.83 \pm 0.13}$	$\mathbf{5.63 \pm 0.14}$	$\mathbf{6.31 \pm 0.06}$	$\mathbf{4.72 \pm 0.10}$

further modified to take low-resolution $3 \times 96 \times 96$ RGB inputs and produce 101 outputs, corresponding to age values from 0 to 100. All models were trained for 100 epochs with a batch size of 64 and optimized based on the widely adopted Adadelta optimization algorithm [272] with a learning rate set to 10^{-1}. To calculate the estimation errors, we adopted the mean absolute error measure (MAE, Eq. 3.2).

5-fold validation. The first set of experiments was based on 5-fold validation and the most common age estimation method (i.e., the MSE regression method). The datasets were randomly divided into 5 equally-sized parts, and for each of the 5 folds, a different part was used for validation, while the rest of the data was used for training. This way, 5 different 80-20 training-validation splits were created and used to achieve robust verification based on a mean and standard deviation of the 5 runs. To further show that the proposed filtering method is beneficial even in the case of highly specialized transfer learning, we performed additional fine-tunings on the training part of a separate benchmark dataset. The fine-tunings were performed for 500 epochs with SGD optimization [76] and a relatively low learning rate (10^{-5}). Separate models were trained based on real and apparent age labels.

5-method validation. The second set of experiments was designed to evaluate if consistent performance gains are obtainable for each of the 5 age estimation methods specified in the previous section. This would indicate that the benefits of the proposed filtering are method-invariant. As mentioned in the previous section, an identical CNN backbone model was used for all methods. 80% of the data from the evaluated dataset was used for training and 20% for validation, while testing was performed on a separate unconstrained benchmark dataset.

Evaluation results

The results of the 5-fold evaluation are presented in Figure 5.5 and Table 5.1. Figure 5.5 shows the average training and validation MAEs over 100 training epochs for the 6 different dataset variations. In addition to the benefits of standard data processing, the graphs show that the proposed data filtering introduces a further large reduction in training and validation errors. The results presented in Table 5.1 further support the proposed filtering. Compared to the processed CACD and IMDB data, the respective filtered data 5-fold results were improved by 1.20 and 2.63 years for the validation MAE, 1.78 and 1.68 years for the real age testing MAE, and 1.70 and 1.57 years for the apparent age testing MAE. Even after the specialized finetunings, versions pretrained on the filtered data consistently yielded improved results of up to 0.48 years, regardless of the type of age estimation task. Note that the high CACD testing errors were caused by the lack of young and older people in the CACD dataset, as shown in Figure 5.3, and were greatly reduced after the Appa-Real finetunings. Compared to the baseline raw data, the performance improved by up to 4.31 years.

The results of the 5-method evaluation are presented in Table 5.2. The results demonstrate considerable performance gains regardless of the age estimation method. Compared to the

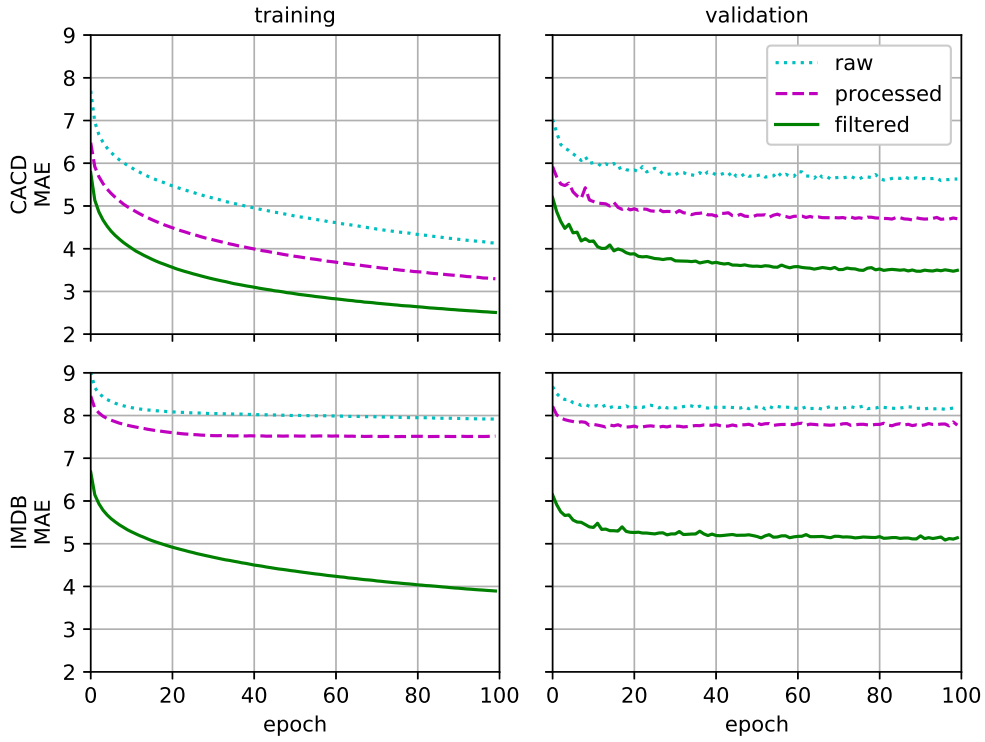

Figure 5.5: Average training and validation MAEs for the first 100 epochs of 5-fold trainings on the raw, processed, and filtered versions of the CACD and IMDB datasets. The first row presents results for the CACD dataset variants (CACD-R, CACD-P, and CACD-F), while the second row presents results for the IMDB dataset variants (IMDB-R, IMDB-P, and IMDB-F).

Table 5.2: Results of the 5-method validation of the raw (R), processed (P), and filtered (F) versions of the CACD and IMDB datasets. The testing was performed on the Appa-Real benchmark. The values represent real age mean absolute error (MAE).

Dataset	Images	Softmax	Regression	DEX [89]	MV [90]	DLDL-v2 [91]
CACD-R	163,446	15.787	14.730	14.513	14.164	14.886
CACD-P	150,383	14.045	13.605	13.420	12.988	13.338
CACD-F	130,571	**12.545**	**11.390**	**11.854**	**11.453**	**11.589**
IMDB-R	460,723	11.547	11.109	11.650	11.160	10.938
IMDB-P	451,571	9.667	8.568	9.338	8.503	8.491
IMDB-F	216,595	**7.807**	**7.080**	**7.349**	**6.771**	**6.787**

processed version, the filtered version of the CACD dataset improved the results by between 1.50 and 2.21 years. In the case of the IMDB dataset, the improvement was similar, ranging between 1.49 and 1.99 years. Compared to the baseline raw data, the performance improved by up to 4.39 years.

For both datasets and both types of age estimation tasks, the robust 5-fold and diverse 5-method validations demonstrated the advantages of using versions of the datasets filtered by the proposed method. Expectedly, the raw versions of the dataset resulted in the worst-performing age estimation models. The standard simple data processing techniques resulted in significant performance gains. Furthermore, the direct subsets of the processed data, obtained by the proposed filtering method, resulted in additional substantial performance gains, thus demonstrating the benefits of the proposed method and highlighting that the data quality is more important than the data quantity. Interestingly, the results also demonstrate that training data processing and filtering can result in a larger margin of improvement than using more advanced age estimation methods.

5.3 Biometrically filtered famous figure dataset

This section presents the designed strategy undertaken to derive a new age estimation dataset, extensive evaluation results, and the comparison with state-of-the-art.

In Section 5.2.3, we demonstrated that the application of the unsupervised biometric filtering method, introduced in Section 5.2.1, produced superior versions of the IMDB and CACD datasets, respectively denoted as IMDB-F and CACD-F. To create a new famous figures dataset with improved properties, we explore options for combining and further filtering the publicly available web-scraped data from IMDB-F, CACD-F, and WIKI datasets.

5.3.1 Dataset design

As the first step in dataset design, we analyze the interrelation between the IMDB-F, CACD-F, and WIKI datasets. Since all three datasets were collected by scraping images of famous figures, and they all provide the subjects' names in the metadata, we decided to analyze the identity overlaps of the three datasets.

For the purpose of identity overlap analysis, we convert the originally provided names from plain Unicode versions to a representation that contains only lower-case ASCII alphabetic symbols. This step helps us to mitigate name-matching failures by eliminating potentially ambiguous white-space, diacritic, and other special symbols. Figure 5.6 visualizes obtained identity overlaps between the three datasets.

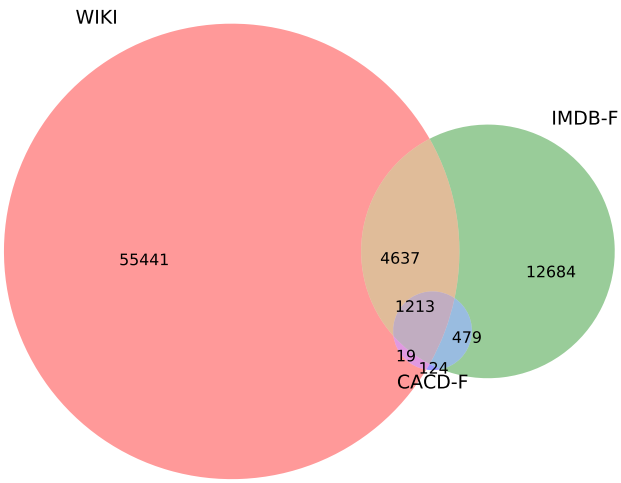

Figure 5.6: Identity overlaps for the WIKI, IMDB-F, and CACD-F datasets.

By combining information from Table 4.1 and Figure 5.6, we observe that the WIKI dataset, although having the lowest number of samples, has the highest number of unique identities. In fact, the WIKI dataset contains only one image per sample. Compared to the WIKI dataset, IMDB and CACD datasets have lower numbers of unique identities but have large average numbers of samples per subject, thus enabling us to apply the proposed filtering method from Section 5.2.1. While IMDB-F and CACD-F datasets can contribute with large amounts of filtered samples and substantial dataset depth, adding samples from the WIKI dataset can potentially increase the overall number of unique identities by 280%, hence greatly improving the dataset breadth.

We apply the initial processing described in Section 5.2.2 to the raw WIKI data to produce the WIKI-P dataset version. To further enhance the WIKI-P dataset, we take advantage of the observation from Section 5.2.2 and Figure 5.4; non-facial images (i.e., false positive detections) tend to form clusters in the face recognition descriptor feature space. We exploit this finding to design a simple false positive detection filtering method.

Finally, leveraging the interrelation between the three datasets, we design an additional majority voting filtering approach that takes advantage of the dataset identity overlaps.

Biometric false positive detection filtering

Once again, we *pick out the bad apples* by designing a biometric clustering approach similar to the method proposed in Section 5.2.1, but applicable to the WIKI data. This simplified version of the aforementioned method does not require multiple images of the same subject, as the formed clusters are not subject-based. The method consists of 3 main steps: feature extraction, sample clustering, and elimination of the *bad* clusters containing false positive detections.

For the feature extraction step, we retain the face recognition descriptor extraction step that provides a compact and highly descriptive numerical representation of facial images with a tendency to cluster non-facial samples, as observed in Section 5.2.2. Whereas we chose to use a graph-based clustering approach to implement the filtering method proposed in Section 5.2.1, the distance-based K-means clustering algorithm is a better fitting choice for this filtering step, as it also provides cluster centers useful in the final cluster elimination step. The K-means clustering is described in more detail in Section 2.1.2. The candidate clusters are formed by applying K-means clustering with an arbitrary set K. Figure 5.7 shows four representative cluster samples obtained on the WIKI-P data with K set to 64.

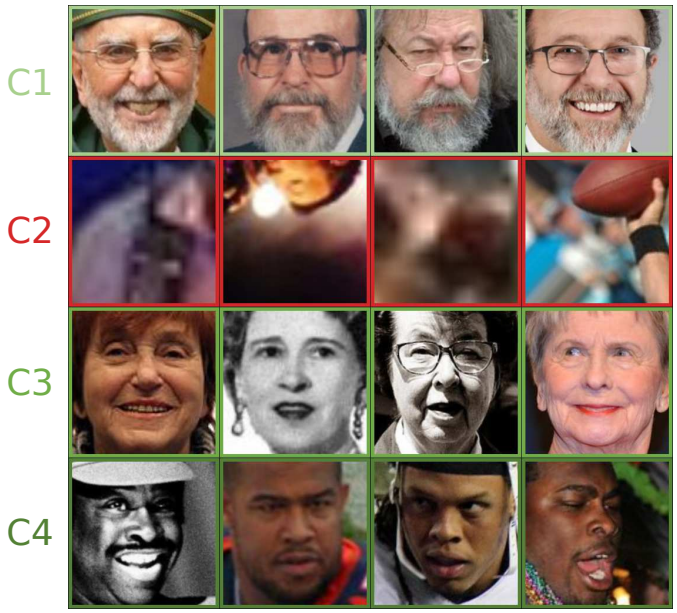

Figure 5.7: Representative WIKI dataset samples from 4 clusters (C1, C2, C3, and C4) obtained by K-means clustering with K set to 64. While clusters C1, C3, and C4 contain similar samples w.r.t. face recognition descriptors, cluster C2 groups mostly non-facial samples.

To identify clusters with false positive detections, we choose a single seed false positive detection sample and compare its face recognition descriptor with the central descriptor of each

of the K-means clusters. The second row in Figure 5.7 shows representative samples from a WIKI-P false positive detection cluster identified by this automatic approach.

This method can be applied to detect non-facial outliers in any type of facial image dataset. We further apply it to CACD-F and IMDB-F datasets to achieve an additional level of data refinement.

Majority vote filtering

To avoid conflicting labels that could emerge in the process of merging the IMDB-F, CACD-F, and WIKI-P data, we leverage the identity overlaps shown in Figure 5.6 to implement a majority vote filtering method that further refines the data.

Once again, we base our filtering method on face recognition descriptors. Additionally, we rely on name-based identifiers, formed as described at the beginning of Section 5.3.1. Algorithm 1 delineates our majority vote filtering approach, while Figure 5.8 shows representative outlier samples identified by the described algorithm.

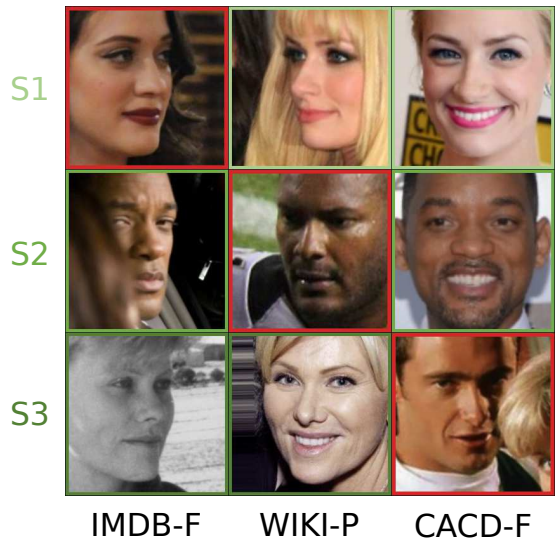

Figure 5.8: Outliers detected by majority vote filtering for subjects S1, S2, and S3 from datasets IMDB-F, WIKI-P, and CACD-F. Outliers are placed on the main diagonal and marked with red frames (best viewed in color).

Assuming that in Algorithm 1 $A, B \in [I, C, W]$ denotes a dataset from a set including IMDB-F, CACD-F and WIKI-P datasets, I_A represents set of identities in the dataset A, S_A^{id} represents set of samples from the dataset A corresponding to the specific identity id, and D_A^{id} represents the mean descriptor for all samples in S_A^{id}. Furthermore, $D_A^{id} \approx D_B^{id}$ denotes that descriptors for identity id from datasets A and B are similar, while $mean(D_A^{id}, D_B^{id})$ represents the mean descriptor for descriptors D_A^{id} and D_B^{id}.

Mean descriptors are calculated by simple averaging. Cosine similarity is used as the similarity measure, with the similarity threshold set to 0.25. This relatively loose threshold, along

Algorithm 1: Majority vote filtering

Input: I_I, I_C, I_W, D_I, D_C, D_W, S_I, S_C, S_W
for $id \in (I_I \cup I_C \cup I_W)$ **do**
  if $id \in (I_I \cap I_C \cap I_W)$ **then**
    if $D_I^{id} \approx D_C^{id}$ ***and*** $D_W^{id} \not\approx mean(D_I^{id}, D_C^{id}))$ **then**
      remove S_W^{id};
    end
    if $D_C^{id} \approx D_W^{id}$ ***and*** $D_I^{id} \not\approx mean(D_C^{id}, D_W^{id}))$ **then**
      remove S_I^{id};
    end
    if $D_I^{id} \approx D_W^{id}$ ***and*** $D_C^{id} \not\approx mean(D_I^{id}, D_W^{id}))$ **then**
      remove S_C^{id};
    end
  else if $id \in (I_I \cap I_C)$ **then**
    if $D_I^{id} \not\approx D_C^{id}$ **then**
      if $|S_I^{id}| < |S_C^{id}|$ **then**
        remove S_I^{id};
      else
        remove S_C^{id};
      end
    end
  else if $id \in (I_W \cap I_C)$ **then**
    if $D_W^{id} \not\approx D_C^{id}$ **then**
      if $|S_W^{id}| < |S_C^{id}|$ **then**
        remove S_W^{id};
      else
        remove S_C^{id};
      end
    end
  else if $id \in (I_W \cap I_I)$ **then**
    if $D_W^{id} \not\approx D_I^{id}$ **then**
      if $|S_W^{id}| < |S_I^{id}|$ **then**
        remove S_W^{id};
      else
        remove S_I^{id};
      end
    end
  end
end

with the mean descriptor comparison approach, reduces the chances of important, difficult facial samples being removed.

5.3.2 Dataset properties

By combining the proposed filtering methods from Sections 5.2.1, 5.3.1, and 5.3.1, and merging the refined IMDB, CACD, and WIKI data, we produce a new derived facial age estimation dataset dubbed Biometrically Filtered Famous Figure Dataset or B3FD in short. The dataset has been made publicly available[¶].

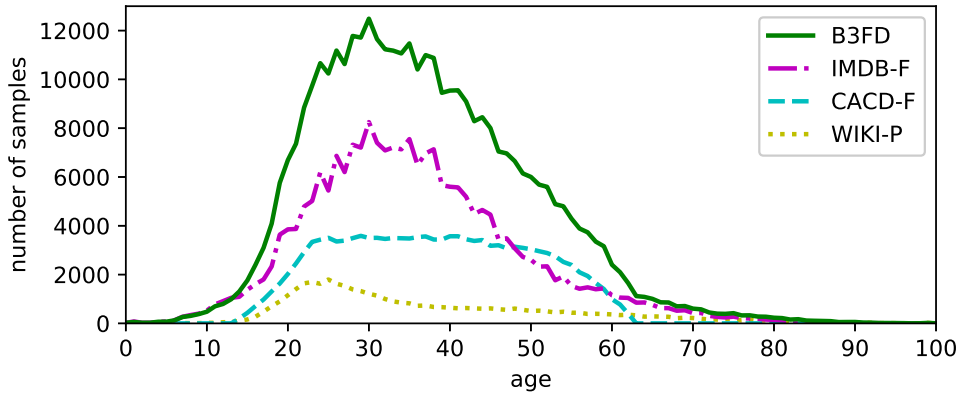

Figure 5.9: Age label distributions for the proposed B3FD dataset and its components (i.e., IMDB-F, CACD-F, and WIKI-P datasets).

B3FD contains 375,592 facial image samples with corresponding age labels. It has 53,759 unique subjects, which amounts to 6.99 samples per subject on average. The age labels are ranging from 0 to 100. The distribution of age labels for the B3FD dataset and its main components (i.e., IMDB-F, CACD-F, and WIKI-P) is shown in Figure 5.9. Our combined filtering efforts resulted in the removal of 310,905 samples, which makes 45.29% of the originally available data.

The B3FD dataset is composed of two main subsets determined by the data origin: the IMDB-WIKI subset (i.e., B3FD-IWS) and the CACD subset (i.e., B3FD-CS). B3FD-IWS consists of 245,204 processed samples from the IMDB-WIKI dataset with 53,568 unique subjects, which amounts to 4.58 samples per subject on average. B3FD-CS consists of 130,388 processed samples from the CACD dataset with 1,831 unique subjects, which amounts to 71.21 samples per subject. These subsets can be useful in case of constraints based on data origin (e.g., only IMDB-WIKI data can be used).

[¶]https://github.com/kbesenic/B3FD

Table 5.3: Architectural details of the used MobileFaceNet-based model. Following the MobileFaceNet notation from [84], rows describe a sequence of operators repeated n times with c output channels. The stride of the first layer in each sequence is denoted by s while t represents the expansion factor applied to the input size.

Input	Operator	t	c	n	s
$\mathbf{128^2 \times 3}$	conv3x3	–	64	1	2
$64^2 \times 64$	depthwise conv3x3	–	64	1	1
$64^2 \times 64$	bottleneck	2	64	5	2
$32^2 \times 64$	bottleneck	4	128	1	2
$16^2 \times 128$	bottleneck	2	128	6	1
$16^2 \times 128$	bottleneck	4	128	1	2
$8^2 \times 128$	bottleneck	2	128	2	1
$8^2 \times 128$	conv1x1	–	512	1	1
$8^2 \times 512$	linear GDConv8x8	–	512	1	1
$1^2 \times 512$	linear conv1x1	–	**101**	1	1

5.3.3 Comparison with the state of the art

This section provides an extensive evaluation of the B3FD dataset and its subsets presented in Section 5.3.2. The evaluation is performed by comparing the performance of models trained on the B3FD dataset variations and the previous state-of-the-art age estimation datasets with respect to real and apparent age estimation accuracy. To demonstrate the generalization capabilities of the models trained on evaluated datasets, the comparison is once again facilitated by a separate unconstrained in-the-wild age estimation benchmark with real and apparent age labels. Furthermore, to show that performance gain is method-invariant, we perform an evaluation based on 5 different age estimation methods, introduced in Section 5.2.3.

Experimental setup

To facilitate this extensive evaluation, we chose to use a CNN architecture from MobileFaceNet family [84] designed specifically for efficient face analysis. Table 5.3 describes the used model architecture in detail. The standard MobileFaceNet architecture was slightly modified to use 128×128 RGB inputs, output 101 values corresponding to ages 0 to 100, and pretrained on the face recognition task. The model has 1M parameters with an inference-time computational cost of 0.99 GFLOPs.

The conventional data processing, explained in Section 5.2.2, was applied to all evaluated datasets. A set of image data augmentations, consisting of randomized horizontal flipping, bounding box perturbations, and in-plane rotations, described in Section 5.2.3, was applied to all training sets, with the only difference being the resulting image resolution (i.e., 128×128).

Table 5.4: Performance of age estimation models trained on different datasets with Softmax and Regression age estimation methods and evaluated on the Appa-Real test set w.r.t. mean absolute error for real and apparent age, along with vote-distribution-based ε-error. R-MAE and A-MAE denote mean absolute errors for real and apparent age, respectively.

		Softmax			Regression		
Dataset	Images	R-MAE	A-MAE	ε-error	R-MAE	A-MAE	ε-error
Appa-Real [30]	7,591	7.252	6.295	0.453	8.287	6.984	0.517
AgeDB [135]	16,488	9.851	9.259	0.545	10.847	9.988	0.563
MegaAge [136]	41,941	9.588	8.434	0.587	10.404	9.364	0.633
MORPH [95]	55,134	13.787	12.478	0.667	13.467	11.698	0.642
CACD [139]	163,446[1]	13.369	12.448	0.631	12.667	11.707	0.636
IMDB-WIKI [89]	523,051[2]	6.509	6.393	0.434	7.294	6.323	0.432
IMDB-WIKI + CACD [89, 139]	686,497[3]	6.986	6.981	0.449	7.819	6.921	0.449
B3FD-CS [ours]	130,388	11.143	10.164	0.592	11.327	10.239	0.598
B3FD-IWS [ours]	245,204	**5.423**	**5.275**	**0.408**	5.828	5.238	**0.408**
B3FD [ours]	375,592	5.547	5.532	0.415	**5.707**	**5.186**	0.409

[1] The processed subset of the data (P) with 150,383 samples was used as it was shown to be superior to the raw data in Section 5.2.3.
[2] The processed subset of the data (P) with 494,158 samples was used as it was shown to be superior to the raw data in Section 5.2.3
[3] The sum of used samples from IMDB-WIKI and CACD is 644,541 for reasons explained in preceding notes.

All evaluated models were trained for 50 epochs with a batch size of 64 and SGD optimization [76]. We set the weight decay parameter to 10^{-4} and the momentum to 0.9. A learning rate scheduler was used to reduce the learning rate by a factor of 10 every 15 epochs. The initial learning rate was set to 10^{-3}.

A random 90-10 training-validation split was used for all datasets except Appa-Real, where the predefined validation set was used, and MegaAge, where the predefined test set was used as the validation set. The models used for dataset evaluation were chosen based on the validation set MAE. All models were evaluated on the Appa-Real test set, as it is the only age estimation dataset providing real and apparent age labels for unconstrained in-the-wild facial images.

We evaluate if the performance gains are indifferent to the type of age estimation task by providing MAE values for both real and apparent age. To additionally evaluate if the performance gains are indifferent to the type of evaluation metric, we also report ε-error [29] based on the distribution of apparent age votes. Formal definitions of these metrics are available in Section 3.2.

Table 5.5: Performance of age estimation models trained on different datasets with three advanced age estimation methods and evaluated on the Appa-Real test set w.r.t. mean absolute error for real and apparent age, along with vote-distribution-based ε-error. R-MAE and A-MAE denote mean absolute errors for real and apparent age estimation, respectively.

		DEX [89]			Mean-Variance [90]			DLDL-v2 [91]		
Dataset	Images	R-MAE	A-MAE	ε-error	R-MAE	A-MAE	ε-error	R-MAE	A-MAE	ε-error
Appa-Real [30]	7,591	7.511	6.382	0.478	7.462	6.355	0.477	7.369	6.099	0.457
AgeDB [135]	16,488	10.023	9.303	0.536	10.378	9.625	0.548	9.851	9.251	0.549
MegaAge [136]	41,941	10.666	9.685	0.643	9.941	8.938	0.623	10.021	9.073	0.633
MORPH [95]	55,134	13.608	12.171	0.652	13.434	11.768	0.642	12.871	11.211	0.614
CACD [139]	163,446[1]	13.538	12.454	0.637	12.555	11.409	0.613	12.079	11.005	0.605
IMDB-WIKI [89]	523,051[2]	7.155	6.673	0.433	7.077	6.259	0.431	6.568	6.101	0.418
IMDB-WIKI + CACD [89, 139]	686,497[3]	7.658	7.174	0.447	7.606	6.831	0.449	7.106	6.704	0.417
B3FD-CS [ours]	130,388	10.914	9.816	0.583	11.033	9.992	0.587	10.844	9.869	0.590
B3FD-IWS [ours]	245,204	5.340	**5.072**	**0.388**	**5.394**	**4.808**	**0.386**	5.158	**4.684**	**0.383**
B3FD [ours]	375,592	**5.282**	5.118	0.393	5.441	5.025	0.403	**5.077**	5.064	0.408

[1] The processed subset of the data (P) with 150,383 samples was used as it was shown to be superior to the raw data in Section 5.2.3.
[2] The processed subset of the data (P) with 494,158 samples was used as it was shown to be superior to the raw data in Section 5.2.3
[3] The sum of used samples from IMDB-WIKI and CACD is 644,541 for reasons explained in preceding notes.

Evaluation results

The results of the experimental evaluation of the proposed B3FD dataset are presented in tables 5.4 and 5.5. Table 5.4 presents evaluation results for the two most frequently used age estimation methods: the classification-based Softmax method and the MSE regression method. Table 5.5 presents results for three additional advanced age estimation methods described in Section 2.3: DEX, Mean-Variance, and DLDL-v2.

To validate if the proposed B3FD dataset outperforms its main components, we evaluated the CACD and IMDB-WIKI datasets under identical conditions. Furthermore, to validate the proposed merging strategy from Section 5.3.1, we also evaluated the performance of the naively merged IMDB-WIKI + CACD dataset. To validate if B3FD provides better generalization capabilities than manually collected data, we evaluated the performance of the largest manually collected dataset (i.e., MORPH) and the more recent in-the-wild AgeDB dataset. To validate if B3FD outperforms the largest in-the-wild dataset for apparent age estimation, we evaluated the performance of the MegaAge dataset. Finally, to validate if the proposed dataset outperforms manually collected domain-specific data, we evaluated models trained on the training part of the evaluation Appa-Real benchmark.

The results show that the models based on the proposed B3FD dataset outperformed models trained on the other evaluated datasets by a notable margin, without exception. The experiments demonstrated that the performance is improved regardless of the used age estimation method, type of age estimation task, or metric. Compared to the naive combination of all evaluated web-scraped data, our derived B3FD dataset provided MAE reduction between 1.44 and 2.38 years for real age and between 1.45 and 2.06 years for apparent age. Compared to the MORPH dataset, the MAE reductions ranged from 6.15 to more than 8 years, most probably due to the constraints and biases associated with manual data collection. The most recent in-the-wild manually collected AgeDB dataset outperformed the larger manually collected MORPH dataset but still fell behind the proposed B3FD data by between 4.30 and 5.14 years for real age and 3.73 and 4.82 years for apparent age. The proposed B3FD dataset also outperformed the MegaAge dataset by a similar margin for both real and, more importantly, apparent age, despite MegaAge providing apparent age labels, while B3FD provides real age labels. Even in the case of the domain-specific manually collected in-the-wild data from Appa-Real, the performance was improved by a considerable margin of up to 2.58 years. B3FD-CS outperformed its corresponding CACD superset by up to 2.62 years for real and 2.64 for apparent age, while B3FD-IWS outperformed its corresponding IMDB-WIKI superset by up to 1.81 years for real and 1.60 years for apparent age. B3FD-IWS even partially outperformed its B3FD superset, presumably due to the age distribution and identity number constraints of the CACD data.

Once again, the results highlighted that the training data quality is more important than the training data quantity and that the performance gains obtained by applying more advanced age

estimation methods can be fairly less significant than the performance gains obtained by utilizing more adequate and filtered training data. This becomes more obvious in our experimental setup, where all age estimation methods are based on the same feature extraction backbone model. For example, the difference between the worst and the best-performing age estimation method on the combined IMDB-WIKI and CACD data is 0.83 years, while using the corresponding filtered B3FD dataset reduces the real age MAE by up to 2.38 years.

5.4 Discussion

Compared to the manually collected facial datasets for biometric trait estimation, datasets collected by automatic web-scraping methods are far superior regarding the sample count and variety, but have a significant downside in terms of label noise. The filtering methods for label noise reduction often require dataset-specific trainings and manual efforts.

The proposed method for unsupervised biometric data filtering can automatically reduce the number of erroneous samples in facial web-scraped datasets by combining only a few general-purpose algorithms. The implemented filtering pipeline resulted in strong sample count reduction on two state-of-the-art web-scraped facial datasets as up to 52% of samples were discarded. The robust 5-fold and diverse 5-method validations with cross-dataset testing both demonstrated that the models based on the filtered data outperform the models based on raw and conventionally processed data by a considerable margin, indicating lower amounts of faulty samples and improved label consistency. The testing of generalization capabilities on the in-the-wild data also indicated that the data diversity is not impaired by the proposed filtering method, despite the strong sample count reduction.

The results obtained by the proposed filtering method were extended by an additional biometric filtering strategy devised to reinforce and refine the merging process of the publicly available web-scraped datasets. The proposed merging process of three different web-scraped data sources resulted in a newly derived, in-the-wild age estimation dataset. The introduced Biometrically Filtered Famous Figure Dataset (B3FD) was experimentally evaluated and compared to both manually collected and web-scraped state-of-the-art datasets. The B3FD dataset consistently outperformed all the evaluated datasets with respect to both real and apparent in-the-wild age estimation. B3FD has been made publicly available.

The extensive experimental evaluation also highlighted the importance of training data quality and label consistency, as the results of models trained on the dataset subsets produced by the proposed filtering methods were superior to the results of models trained on larger datasets, as well as to models trained with more advanced age estimation methods.

Chapter 6

Towards video-based age estimation

Humans and ML models can estimate chronological age from a single image with comparable accuracy. However, when humans are not confident in their estimation, they tend to *take a better look* by examining the subject for a longer period of time and from different viewpoints. This observation is supported by Hadid [251], stating that psychological and neural studies [273, 274, 275, 276] indicate that head-pose and facial expression changes provide important cues for face analysis. Existing image-based methods can be applied directly to video frames. However, Ji et al. [257] point out that deploying image-based age estimation models directly to videos leads to estimation stability issues, while Hadid [251] states that the image-based approach exploits the abundance of frames in videos but ignores useful temporal correlations and facial dynamics. To explore the potential of videos in the age estimation field in more detail, we require a sufficient amount of appropriate video data and a method that is able to leverage video-specific information.

A recurring topic in the reviewed work from Section 4.4 is the lack of publicly available video data with age annotations, suitable for model training and evaluation. For the development and testing of almost every reviewed method, the authors needed to resort to data collection. However, the collected data was not made publicly available in [251, 257, 258]. Moreover, manual data collection often results in very small datasets. Ji et al. [257] proposed a single-subject, single-video dataset, and used it both for attention module training and for video-based testing, making their conclusions questionable. Similarly, Zhang et al. [258] used a two-subject, two-video dataset for their model evaluation. Dibeklioğlu et al. [144] based their initial work on a larger video dataset but used it for both training and evaluation. A very specific nature of the used data, where every subject transitions from neutral to smiling facial expression, potentially caused their model to overfit this specific type of data. Even when they introduced a dataset related to a different facial expression (disgust, in [145]), they did not perform cross-dataset testing to verify the generalization capabilities of their models. Pei et al. [256] also followed their evaluation protocol without performing a cross-dataset evaluation. Ji et al. [257], Han et

al. [146], and Zhang et al. [258] did not perform a comparison with previous video-based age estimation methods. The lack of a large, in-the-wild, video-based age estimation benchmark undermines the credibility of some of the reviewed findings, as well as the ability to fairly compare different algorithms and methods. Moreover, the lack of video training data presents a significant hurdle in the development of a method that truly harnesses video information.

This chapter is based on our work from [277]* that *takes a better look* at the potential of videos and video-based methods to refine automatic facial age estimation. We focus on resolving two prominent issues in this field of research: the lack of a well-defined, publicly accessible, benchmark and the inability to train video-based models due to an absence of labeled training data. The main contributions of this work are summarized as follows:

- We reproduce the recent frame-based age estimation benchmark from [147], achieve state-of-the-art results on their protocol, and make the missing benchmark metadata publicly available.
- We design a new, video-based age estimation benchmark that utilizes data from [147], extend the metadata by using a commercial face tracking system, perform an exploratory analysis, and make the benchmark protocol, along with all the required metadata and video processing frameworks, publicly available.
- We design a semi-supervised video age estimation method that overcomes the lack of labeled training data and outperforms its image-based counterpart, thus setting baseline results on the proposed benchmark.

The rest of the chapter is organized as follows. As it is a prerequisite for any insightful and fairly comparable research, we first tackle the design of the video age estimation benchmark dataset and protocol in Sections 6.1 and 6.2, respectively. Section 6.3 describes our path towards a video-based age estimation method. First, it explores the potential of videos and some conventional video-based methods. It continues by describing the generation of pseudo-labeled data, the design of a semi-supervised video method, and the obtained benchmark results. Finally, our conclusions are discussed in Section 6.4.

6.1 Video age estimation benchmark data

As reviewed in Section 4.1.2, Casual Conversations are the only publicly accessible video-based datasets suitable for facial age estimation. Moreover, they were curated with a special focus on ethical data collection and demographical fairness. They are also the largest video datasets with precise age annotations, and their licenses allow for the evaluation of both academic and commercial models†. All this makes them great candidates for a video age estimation benchmark.

*Reproduced with permission from SciTePress.

†For specific conditions, please refer to the CC license agreements.

Figure 6.1: CCMiniVID video processing framework. All faces are tracked with a commercial tracking system. The metadata of multi-subject videos is manually filtered to contain only data related to the subject of interest. The framework extracts aligned face crop sequences using the raw CC videos and the produced tracking data.

The authors of the CC dataset explored its potential for age estimation evaluation by performing a set of baseline experiments. For that purpose, they used CCMini; a well-balanced subset of the CC dataset. For further reference, we dub this image set as CCMiniIMG. Although this initial effort towards a CC-based age estimation benchmark motivated our work, we continue by discussing the adequacy of the proposed evaluation protocol for video-based age estimation.

6.1.1 CCMiniIMG

The CCMiniIMG benchmark was made available in the form of pre-extracted raw video frames, where 100 frames are provided for each of the 6,022 videos. However, the authors did not provide face detection metadata or clear information on the frame sampling procedure. The lack of face detection metadata prevents the exact reproduction of their evaluation protocol, as some frames contain multiple subjects. The lack of information on the frame sampling method and the fact that the protocol relies only on frames rather than continuous sequences prevents the evaluation of video-based age estimation methods that leverage temporal information.

6.1.2 CCMiniVID

To design a video benchmark dataset for age estimation based on the CC data, we follow CCMiniIMG and select 6,022 well-balanced videos from the CCMini subset. All videos were recorded at 30 FPS, and the mean video duration is 64.44 ± 13.56 seconds, with 99% of videos lasting for at least 20 seconds. We set our target to extract sequences with 20 seconds of continuous face presence from each video, where possible. We utilize a commercial face tracking

system from Visage Technologies[‡] to produce high-quality frame-level tracking data comprising 75 facial landmark points, apparent pitch, yaw, and roll head angles, tracking quality, and face scale. We proceed with semi-automatic tracking data analysis, as shown in Figure 6.1.

The CC data does not provide correspondence between subjects and labels for the multi-subject videos, potentially causing erroneous subject-level labels. We perform automatic detection of multi-subject video candidates, followed by manual verification and selection of tracking streams in 108 videos. In two of the videos, subjects of interest face away from the camera, making their faces fully self-occluded. These videos are not suitable for a face-oriented evaluation benchmark. Continuous face tracking with a target duration of 20 seconds was achieved in 5,932 of 6,022 videos. Manual verification of the remaining 90 videos showed that in 21 videos, continuous tracking was not sustainable due to occlusions, self-occlusions, camera issues, or extreme lighting. 69 videos were shorter than 20 seconds, with the shortest one lasting for only 6 seconds.

Based on the obtained tracking data and manual verification, we propose three versions of the CCMiniVID benchmark data. CCMiniVID-O, where O stands for *original*, comprises only videos from the original CCMini subset. It contains 5,932 sequences with continuously tracked 600 frames (20s@30FPS) and 90 outlier videos with between 100 and 599 continuously tracked frames. It also includes the aforementioned two videos where subjects of interest are not visible. In CCMiniVID-A, where A stands for *alternatives*, replacements for 92 problematic videos were manually selected from the full CC dataset by looking for the most similar substitutes in video galleries of target subjects. This allows consistent evaluation with 600 consecutively tracked frames for all 3,011 subjects from the CC dataset. Additionally, we propose CCMiniVID-R, where R stands for *reduced*; a simple subset of CCMiniVID-O where the 92 problematic videos were removed. This allows for precise and consistent testing without the need to download the complete 6.9 TB Casual Conversations dataset.

CCMiniVID exploratory analysis

All three versions of the CCMiniVID data are based on videos from the same 3,011 subjects, with two videos per subject selected from CC to create a balanced subset. Videos of 62 subjects that did not provide age information are unsuitable for an age estimation benchmark. That leaves us with 5,898 videos of 2,949 unique subjects used in CCMiniVID benchmark datasets.

The original CC metadata provides some subject-level labels (i.e., age, gender, and skin tone) and video-level labels for lighting (i.e., bright or dark). The authors encourage users to extend the annotations of their dataset [147]. To this extent, we processed the dataset with a commercial face tracking system and filtered the results by automatically processing the tracking data and manually validating edge cases. The tracking data consists of 75 facial landmark

[‡]https://visagetechnologies.com/facetrack/

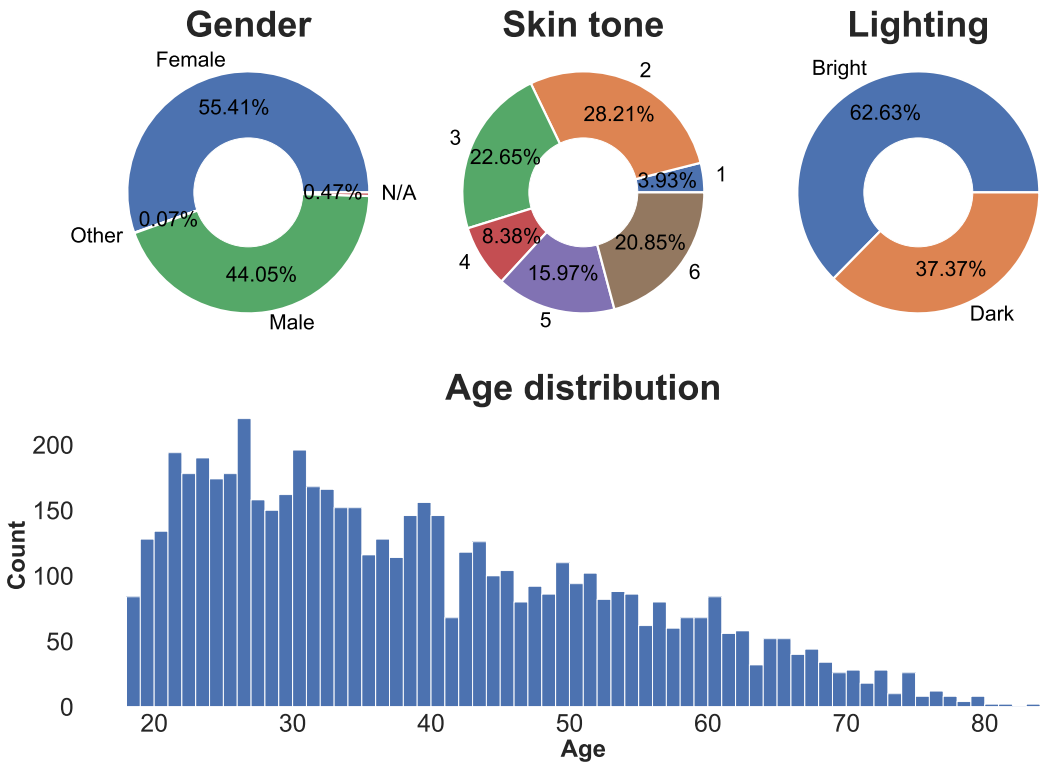

Figure 6.2: CCMiniVID-O label distributions for gender, age, and skin tone (subject-level) and lighting (video-level).

points, apparent pitch, yaw, and roll head rotation, and tracking quality for each frame.

Figure 6.2 presents label distributions for the CCMiniVID-O benchmark dataset based on labels from CC's original metadata. We can see that gender and lighting label distributions are fairly balanced, with some reasonable underrepresentations in the skin tone distribution. CCMiniVID-A shares the same distribution for the subject-level labels (i.e., age, gender, and skin tone). Still, there is a minor difference in the lighting distribution caused by the selection of alternative videos (i.e., 62.67% bright and 37.33% dark). Videos of 10 subjects were dropped in the reduced CCMiniVID-R version, causing a negligible variation in all distributions.

Figures 6.3, 6.4, and 6.5 present distributions for relevant tracking data from our extended CCMiniVID-O metadata. Head-pose angle distributions show that this conversational dataset is oriented towards frontal and near-frontal faces, with some occurrences of extreme head poses. The face scale distribution, where the face scale is the width or height of a square face bounding box in pixels, shows that the provided high-quality videos contain faces mainly in the 100 to 400-pixel range. The tracking quality distribution indicates that continuous face tracking with high confidence was sustained on a large majority of the videos.

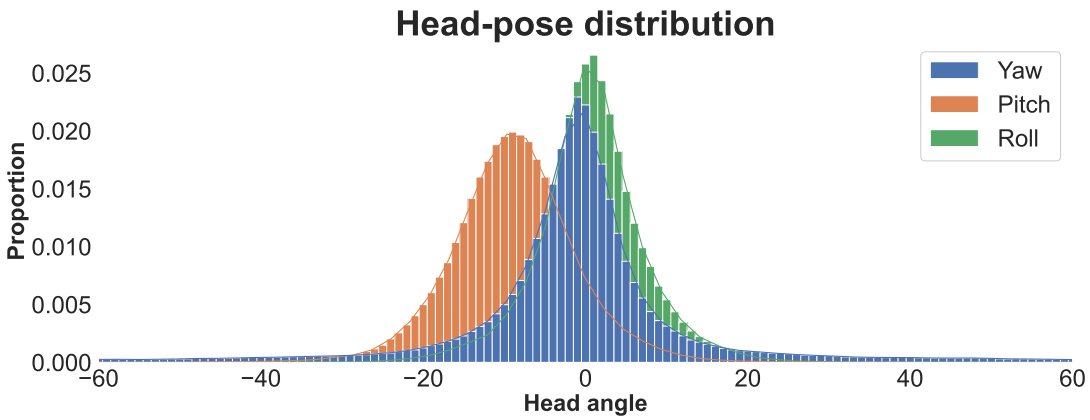

Figure 6.3: CCMiniVID-O head-pose angle distributions, based on the produced face tracking data.

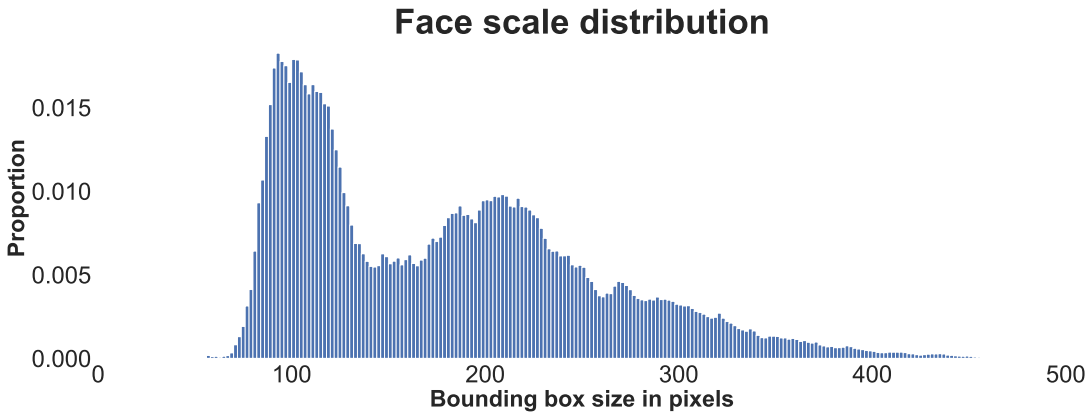

Figure 6.4: CCMiniVID-O face scale distribution, based on the produced face tracking data.

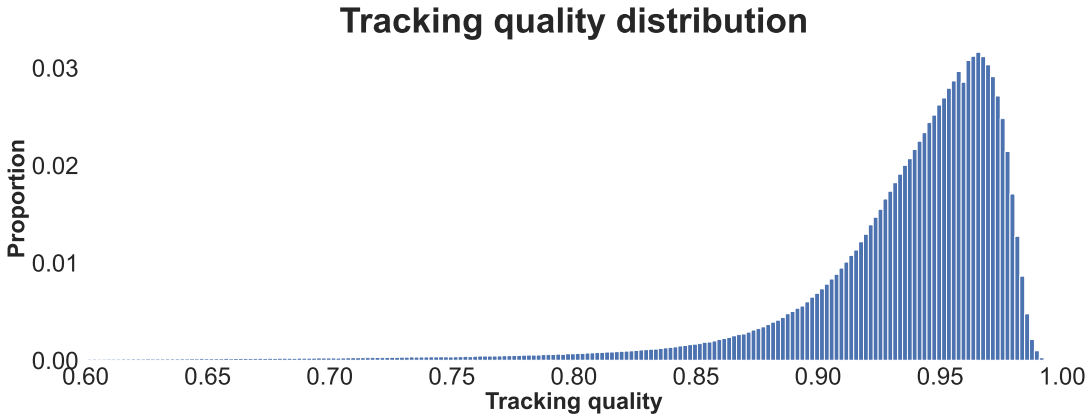

Figure 6.5: CCMiniVID-O tracking quality distribution, based on the produced face tracking data.

6.2 Video age estimation benchmark protocol

To design the benchmark protocol, we first review the image-based age classification protocol introduced in [147].

6.2.1 Image-based benchmark protocol

The authors of [147] extracted 100 frames from each of the 6,022 CCMiniIMG videos and provided only video-level metadata and raw frames. Face detection metadata was not provided, and neither face detection setup nor frame sampling algorithm were specified. The proposed protocol treats age estimation as a 3-class age classification task. The median of 100 image-level age estimations is used as the video-level estimation, which is then mapped to 3 predefined age groups (i.e., 18-30, 31-45, and 46-85). Classification accuracy was selected as the primary metric. The protocol additionally relies on the dataset's auxiliary labels to calculate accuracy across three gender groups, six apparent skin tone types, and two ambient lighting types.

Protocol reproduction details

As discussed in Section 6.1.1, the authors of [147] extracted 100 frames from each of the 6,022 videos and detected faces with the DLIB face detector [278]. Face detection metadata was not provided, and neither face detection setup nor frame sampling algorithm were specified.

To reproduce the evaluation protocol, we first process CCMiniIMG raw frames following the limited information available in [147] and explore DLIB's face detection options. DLIB offers two types of face detectors: HOG-based and CNN-based. The HOG-based face detector is very light, but it is not able to detect many of the challenging samples from this dataset. Therefore, we choose the CNN-based detector. We experiment with DLIB's sole face detection parameter (i.e., upsampling factor) and CLAHE image enhancement [279] with different tile sizes and clip limits to push the detection rate to 99.82%. No protocol for handling multiple detections was specified, so in each frame, we selected the detection with the highest detection confidence. 62 of the 3,011 subjects did not provide age labels, eliminating 12,400 images from the test set. This results in 588,720 video frames with successfully detected faces and valid age labels. We proceeded by extracting face crops using DLIB's affine alignment algorithm, which uses five facial landmark points detected by DLIB's shape predictor. Face crops were extracted with 50% context and resized to 256×256p.

To enable easy reproduction of this evaluation protocol, we make DLIB's detection metadata and CCMiniIMG frame processing scripts publicly available[§] and encourage authors to utilize it in their work to avoid inconsistencies.

[§]https://github.com/kbesenic/CCMiniVID

Table 6.1: Age classification accuracy results for the baseline methods from [147] and our baseline image-based age estimation model on the CCMiniIMG benchmark data according to the image-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
[185]	38.05	37.44	39.48	66.67	39.56	38.72	40.84	36.47	36.47	34.89	38.49	37.04
[280]	42.26	42.28	44.53	100.00	42.33	41.78	42.30	42.79	42.44	37.99	42.94	41.12
[281]	54.32	54.21	56.18	83.33	46.51	55.52	54.59	55.78	53.78	52.57	54.17	55.20
[267]	73.06	70.20	76.48	87.50	76.72	82.15	72.90	70.24	67.52	65.61	73.90	71.64

Table 6.2: Age classification accuracy results for our baseline image-based age estimation model on the three versions of CCMiniVID benchmark data according to the image-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
CCMiniIMG	73.06	70.20	76.48	87.50	76.72	82.15	72.90	70.24	67.52	65.61	73.90	71.64
CCMiniVID-O	73.60	70.90	76.75	93.75	77.59	81.31	74.25	74.09	65.82	67.48	75.18	70.96
CCMiniVID-A	73.50	70.56	77.02	87.50	77.59	81.55	73.13	74.09	66.67	67.24	75.00	70.98
CCMiniVID-R	73.59	70.85	76.81	93.55	77.23	81.38	74.43	73.88	65.70	67.49	75.17	70.99

Image-based baseline results

We evaluate the performance of our best-performing image-based age estimation model from [267] and Section 5.3.3 and compare it to the previously reported SOTA results from [147]. The selected model is based on a CNN architecture from MobileFaceNet family [84] (Section 5.3.3) and trained on the proposed large in-the-wild B3FD image dataset (Section 5.3.2) using the DLDL-v2 age estimation method [91]. The model takes 128×128 RGB face crop inputs and outputs 101 values corresponding to ages 0 to 100. More details are available in Sections 2.3.3 and 5.3.3. In this chapter, we call this model the *baseline image-based age estimation model*.

Table 6.1 presents the baseline results. The selected model outperforms the overall age classification accuracy of the previously reported leading method from [281] by a large margin of 18.74 points, establishing the new state-of-the-art on this benchmark. Although it outperforms the baseline methods by a large margin, it shows a similar relative performance drop for female subjects, darker skin tones, and poorly illuminated recordings.

To validate the CCMiniVID benchmark data from Section 6.1, we apply the image-level baseline model and protocol to its three versions. Table 6.2 compares results obtained on the original CCMiniIMG data. The results are closely matched, showing that while the CCMiniVID benchmark data offers several benefits over CCMiniIMG, such as continuous video sequences, face tracking data, and clear mapping between video subjects and the dataset's metadata, it retains a very similar difficulty level.

6.2.2 Video-based benchmark protocol

Methods that work with sequential (i.e., temporal) data can be either offline or online. Offline methods process a video as a unit non-causally and produce a single estimate for the whole video. Temporal stability is not an issue since only one estimate per video is produced. Therefore, we find Mean Absolute Error (*MAE*, Eq. 3.2) a sufficient metric for offline estimation methods. The absolute error is calculated based on a single estimate per video, while the mean is calculated over all videos in the dataset. Online methods produce updated age estimations in real time as new video frames are captured. For online methods, we propose Temporal Mean Absolute Error (*tMAE*) and Temporal Standard Deviation (*tSTD*) as the benchmark metrics to evaluate both method's accuracy and online estimation stability, respectively. These are standard *MAE* and *STD* metrics calculated at the frame level, as online methods produce new estimates for each of the frames in the temporal dimension. Following the CCMiniIMG protocol, we rely on the dataset's additional labels to calculate the proposed metrics across three gender groups, six apparent skin tone types, and two ambient lighting types. To unambiguously define the benchmark protocol, we make the metadata (including the face tracking data from the

Table 6.3: Offline age estimation *MAE* for our baseline image-based age estimation model on the three versions of CCMiniVID benchmark data according to the offline video-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
CCMiniVID-O	5.25	5.61	4.81	3.82	5.14	4.78	5.02	5.15	5.56	5.95	5.04	5.60
CCMiniVID-A	5.25	5.64	4.78	3.63	5.03	4.74	5.05	5.14	5.59	6.00	5.00	5.67
CCMiniVID-R	5.23	5.60	4.78	3.89	4.94	4.79	4.97	5.19	5.57	5.92	5.01	5.59

Table 6.4: Online age estimation *tMAE* and *tSTD* for our baseline image-based age estimation model on the three versions of CCMiniVID benchmark data according to the online video-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
CCMiniVID-O	6.19±3.25	6.77±3.73	5.47±2.66	4.77±2.61	6.19±3.44	5.64±2.98	5.84±2.99	5.98±3.01	6.63±3.60	7.04±3.71	5.97±3.13	6.56±3.46
CCMiniVID-A	6.10±3.12	6.68±3.58	5.39±2.54	4.43±2.54	6.03±3.33	5.53±2.84	5.78±2.88	5.84±2.84	6.58±3.44	6.98±3.57	5.85±3.02	6.52±3.29
CCMiniVID-R	6.17±3.25	6.75±3.73	5.45±2.66	4.83±2.61	6.02±3.44	5.64±2.98	5.79±2.98	6.01±3.02	6.64±3.60	7.01±3.71	5.94±3.13	6.54±3.46

commercial tracking system), the test set extraction framework, and the evaluation framework publicly available athttps://github.com/kbesenic/CCMiniVID.

Video-based baseline results

Tables 6.3 and 6.4 present results of our image-based age estimation model described in Section 6.2.1 on the three CCMiniVID benchmark dataset versions, following the previously described offline and online video protocols, respectively. To comply with the offline estimation protocol, which expects a single estimate per video, we use the median of the frame-level estimations to calculate MAE. The online protocol's $tMAE$ and $tSTD$ metrics are calculated directly from the frame-level estimation errors.

The results obtained on the three versions are very similar, and the relative performance drop can be observed in the case of female subjects, darker skin tones, and lack of good lighting. High $tSTD$ numbers indicate unstable age estimation across video frames. The large gap between overall offline and online MAE indicates that video-level estimations are more precise than frame-level estimations.

6.3 Towards video-based age estimation method

All previously presented results are obtained with image-based age estimation methods. The central premise of this work is that age can be estimated more precisely by *taking a better look*, i.e., by using a longer video sequence instead of a single image. To verify this, we first analyze how video sequence duration affects the accuracy of age estimation.

6.3.1 Taking a better look

For this experiment, we use the CCMiniVID-A video benchmark data, which contains 600 frames (20s@30FPS) of continuous face tracking for two videos of each subject in the CCMini dataset. We find this version of the benchmark data most suitable since it contains all subjects, and all video subsequences are of the exact same duration. The majority of video-based methods rely on frame subsampling techniques to avoid redundant processing of nearly identical neighboring frames and to reduce computational complexity. We set the subsampling step to 6, following various video processing methods [282, 283, 284, 285, 286]. We calculate age estimation MAE for video subsequences lasting from a single frame (0 seconds) to 20 seconds. The results are presented in Figure 6.6, denoted as 2D CNN.

By using video sequences of 1 second, the single-frame-based estimation error is reduced by 5.59%. By using 5 seconds, the error is reduced by 10.29%. Increasing sequence duration from 5 to 15 seconds reduces the error by an additional 2.67 percentage points. However, increas-

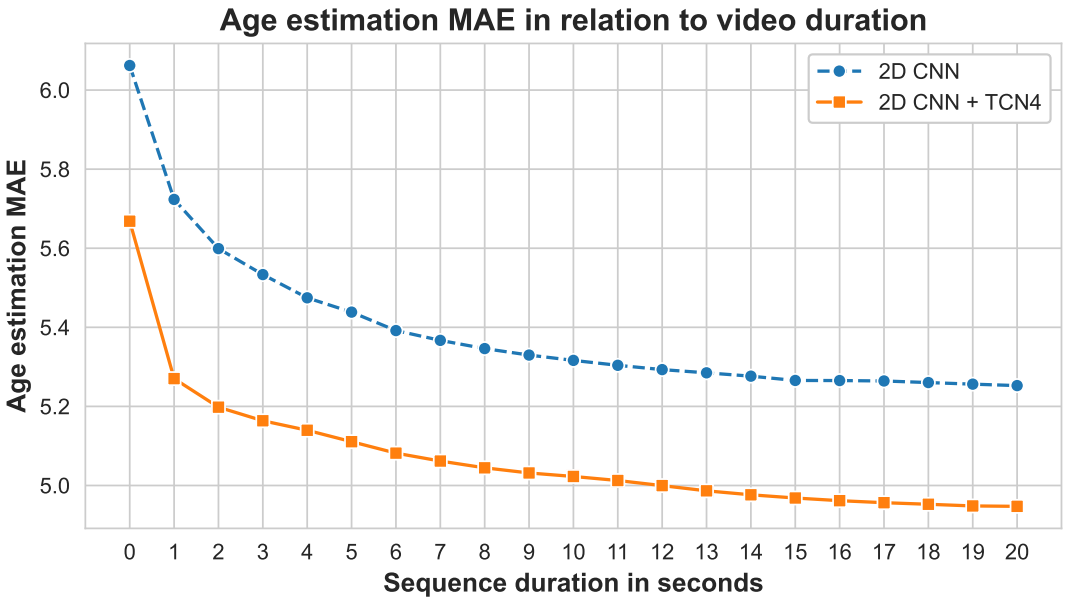

Figure 6.6: Age estimation *MAE* of our baseline 2D CNN image-based age estimation model and the proposed video-based model (2D CNN + TCN4) in relation to video sequence duration on the CCMiniVID-A offline video-based benchmark protocol.

ing the sequence duration from 15 to 20 seconds results in almost no improvement. Using the complete sequences reduces the estimation error by 13.35%, compared to the single-frame approach. The results demonstrate that *taking a better look* is very useful for age estimation, even with a basic image-based estimation method and simple median aggregation of the frame-level results.

6.3.2 Cherry-picking based on tracking data

Cherry-picking is a popular name for a technique based on the selection of the best or most suitable elements from a set. A frontal-face cherry-picking approach was proposed by [258], where a fixed threshold was used on the sum of absolute head-pose angles. To evaluate the cherry-picking approach, we can utilize the produced frame-level tracking data described in Section 6.1.2 (i.e., head-pose angles, tracking quality, and face scale). Using a fixed threshold is not suitable for our evaluation protocol since it might cause biases in video-based or even subject-based label distributions. For example, there might be more female subjects with bad lighting in the non-frontal subset. To mitigate this issue, we extract three equally sized subsets from each of the benchmark videos. As a baseline, we extract the chronologically first 50% of the video frames. We also extract the best and the worst 50% of the video frames based on a certain criterion. This ensures that all videos and subjects are used in all three subsets and that the subsets are of equal size. Results of this experiment are presented in Table 6.5.

Contrary to the findings in [258], head-pose angles seem to have negligible influence on

Table 6.5: Age estimation *MAE* of our baseline image-based age estimation model on the CCMiniVID-A benchmark data for frame cherry-picking approaches based on different face tracking data.

Criterion	Yaw	Pitch	Roll	Tracking quality	Face scale
First 50%	5.29	5.29	5.29	5.29	5.29
Best 50%	5.32	5.26	5.29	5.27	5.27
Worst 50%	5.26	5.32	5.26	5.30	5.31

age estimation error in our experiments. The image-based age estimation model was trained on the proposed B3FD dataset, a very large in-the-wild age estimation dataset with unconstrained head poses. The model was also trained with data augmentations to make it robust to tracking instabilities and low face resolution, further explaining why filtering based on auxiliary tracking data did not result in significant improvements.

6.3.3 Video-based age estimation method

The main obstacle to training an age estimation model that leverages facial dynamics and temporal information from videos is the lack of video data with age labels. As mentioned in our data review in Section 4.1.2, there are no publicly accessible video datasets that permit training of age models. As reviewed in Section 4.4, some researchers resorted to a manual collection of very small video datasets (e.g., only one or two subjects). We believe this is insufficient to train a reliable model with good generalization capabilities.

To deal with the lack of annotated video data that restricts the usability of fully supervised learning, we explore pseudo-labeling and semi-supervised learning. Whereas supervised learning requires labels for all training samples, semi-supervised learning algorithms aim at improving their performance by utilizing unlabeled data [287]. According to [55], this sounds like magic, but it can work under the right conditions. One approach for making use of unlabeled data is generating pseudo labels. Pseudo labels are weak labels generated by the model itself and subsequently used to further train the model [288]. Semi-supervised learning is explained in more detail in Section 2.1.3. In our setup, we leverage the fact that the subject's age does not change during a single video recording. We rely on an image-based age estimation model and apply it to every frame of an unlabeled facial video dataset. The frame-level results are aggregated to get video-level pseudo labels. These pseudo labels can then be used to supervise the learning of spatiotemporal models. In the case of end-to-end training, the image-based backbone used to generate the pseudo-labels is also further optimized and improved.

Pseudo-labeled training data

The main advantage of the proposed approach is that any source of unlabeled facial videos can be used for model training. We combine two large sources of raw videos. The video portion of the IJB-C dataset [289] is selected since the dataset's general statistics indicate good age distribution. The CelebV-HQ dataset [290] is chosen for its size and good distribution concerning facial expressions, appearance attributes, and actions. The raw videos were once again processed with a commercial face tracking system, and frame-level estimations were produced with the image-based age estimation model from Section 6.2.1. The median was used to aggregate frame-level estimations into video-level pseudo labels. We filtered the video data with respect to face scale, tracking quality, and sequence duration. The produced set consists of 28,619 videos with a total of 7,535,299 frames, averaging 263 continuously tracked frames per video. Pseudo-label distribution ranges from 6 to 89 years.

Semi-supervised video-based method

In Section 6.2.2, we demonstrated that offline median estimation across video sequences can be much more precise than frame-level predictions. Our goal is to use the median pseudo labels to train a temporal model that will be able to replicate that performance boost, but in an online fashion and based on a much shorter time window.

Our method follows the fundamental design principle of many video-processing methods [283, 291, 292, 293, 294], including some previously reviewed video age estimation methods [256, 257], meaning that the model consists of an image-based 2D CNN feature extractor and a temporal model that learns to aggregate frame-level features in an optimal way.

For feature extraction, we once again rely on our 2D CNN model for image-based age estimation from Section 6.2.1. The feature extractor backbone is stripped of its last age classification layer to produce features of 512 elements. Motivated by the results of [295], we base our temporal model on their implementation of the Temporal Convolutional Network (TCN)[¶]. The proposed TCN can be parameterized with respect to the number of layers, kernel size, and number of input and output channels. We set the kernel size to 2 and parameterize the TCN to input and output feature vectors of size 512, matching the chosen feature extractor output. The TCN's output is passed to a fully connected layer that translates the temporal feature vectors to age estimations. The number of layers determines the network's receptive field and, therefore, the time window size used in online processing. We experiment with 3 and 4 layers, resulting in receptive fields of 8 and 16 frames, respectively.

An overview of the method is presented in Figure 6.7. The figure illustrates the processing pipeline for a 4-layer TCN with a receptive field of 16 frames. The method predicts a softmax

[¶]github.com/locuslab/TCN/blob/master/TCN/tcn.py

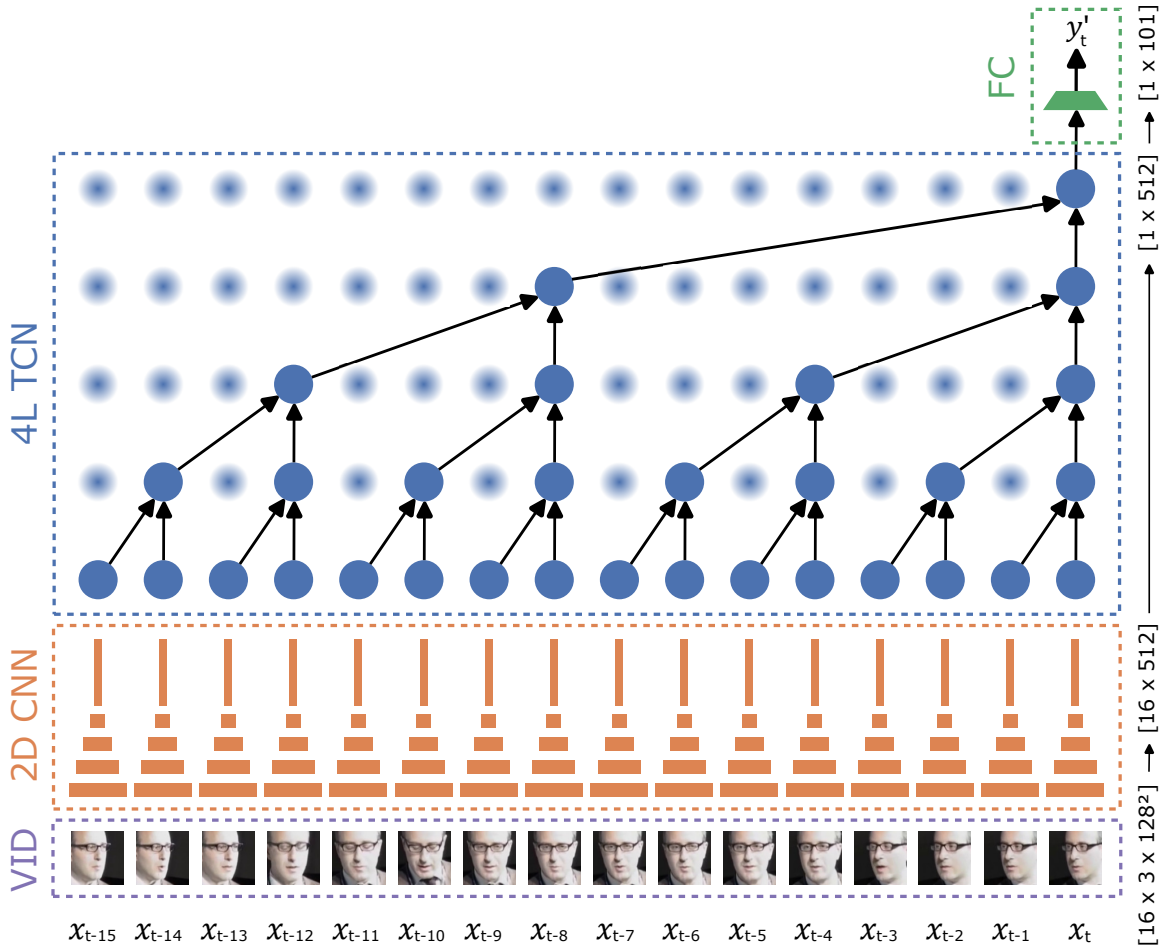

Figure 6.7: The proposed video age estimation method based on a 2D CNN feature extractor and a 4-layer TCN temporal model (4L TCN), followed by a fully connected layer (FC) for dimensionality reduction. The model takes 16 input video frames (VID) denoted as x_t to x_{t-15} to produce an age estimation probability vector y'_t.

probability vector y'_t of size 1×101 (mapping to ages from 0 to 100) based on the current frame face crop x_t and 15 previous face crops (x_{t-1} to x_{t-15}). The weighted sum of the softmax probabilities is used as the final age prediction, according to the DLDL-v2 method.

We train the proposed model in an end-to-end manner, meaning that we jointly optimize both the image-pretrained feature extractor and the randomly initialized TCN. The training data is divided into training and validation subsets with a random 80:20 split. The model is trained with the DLDLv2 age estimation method and Adam optimization [296], using a learning rate of 10^{-6} and weight decay of 10^{-3}. We adopt an early stopping approach, where training is ended when the validation subset *MAE* plateaus.

Video-based benchmark results

The results in Tables 6.6 and 6.7 show that we outperformed image-based baseline according to both offline and online evaluation protocols, even though our model trainings were supervised with pseudo labels generated by that exact baseline model. Figure 6.6 presents consistent im-

Table 6.6: Offline age estimation *MAE* for our baseline 2D CNN image-based age estimation model and the proposed video-based age estimation models (2D CNN + TCN) on the CCMiniVID-A benchmark data according to the offline video-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
2D CNN	5.25	5.64	4.78	3.63	5.03	4.74	5.05	5.14	5.59	6.00	5.00	5.67
2D CNN + TCN3	5.07	5.39	4.68	3.72	4.98	4.64	4.81	5.03	5.40	5.72	4.78	5.57
2D CNN + TCN4	4.95	5.25	4.58	3.81	4.71	4.49	4.76	4.83	5.27	5.61	4.76	5.27

Table 6.7: Online age estimation *tMAE* and *tSTD* for our baseline 2D CNN image-based age estimation model and the proposed video-based age estimation models (2D CNN + TCN) on the CCMiniVID-A benchmark data according to the online video-based evaluation protocol.

		Gender			Skin type						Lighting	
	Overall	Female	Male	Other	Type I	Type II	Type III	Type IV	Type V	Type VI	Bright	Dark
2D CNN	6.10±3.12	6.68±3.58	5.39±2.54	4.43±2.54	6.03±3.33	5.53±2.84	5.78±2.88	5.84±2.84	6.58±3.44	6.98±3.57	5.85±3.02	6.52±3.29
2D CNN + TCN3	5.36±1.79	5.75±2.07	4.88±1.44	4.12±1.70	5.33±1.82	4.91±1.64	5.07±1.68	5.23±1.64	5.73±1.98	6.05±2.03	5.06±1.71	5.85±1.93
2D CNN + TCN4	5.16±1.51	5.52±1.74	4.71±1.23	4.02±1.40	4.97±1.53	4.69±1.40	4.94±1.42	5.00±1.37	5.52±1.67	5.85±1.70	4.96±1.44	5.49±1.62

provements of the 4-layer TCN model (TCN4) on the offline protocol in relation to sequence duration. Significant improvements can also be observed in the online protocol results in Table 6.7, where the TCN4 overall *tMAE* is reduced by 15.41%, while *tSTD* is reduced by an even larger margin of 51.60%.

The TCN3 model uses a time window of size 8, which amounts to just 1.6 seconds of video data under the proposed frame subsampling setting. By using the first 1.6 seconds of each video, TCN3 can achieve *MAE* of 5.32 with a single online inference pass, compared to 5.52 in the case of the aggregated image-based results. Image-based result aggregation on the full 20-second videos gives *MAE* of 5.25. Using a single online TCN4 inference pass on the first 3.2 seconds of each video, we even outperform the full-video results with *MAE* of 5.15, simultaneously achieving 84% shorter estimation time and improved estimation accuracy.

6.4 Discussion

Aligned with our initial premise for this work, *taking a better look* at video sequences significantly improves age estimation compared to the single-image approach. Our evaluation has also confirmed previous findings regarding inconsistent performance with respect to gender, skin tone, and lighting type, highlighting the issue of demographic biases in the machine learning field.

Contrary to the findings of previous works, our experiments have shown that the correlation between age estimation error of contemporary age estimation models and auxiliary face tracking data (i.e., head-pose angles, tracking quality, and face scale) is negligible, making the frame cherry-picking method ineffective.

The proposed video age estimation method, based on pseudo-labeling and semi-supervised learning, overcomes the lack of available annotated training data and improves age estimation accuracy according to both offline and online evaluation protocols, while estimation stability is improved by more than 50%. Our goal is for our carefully designed and publicly available benchmark protocol, along with the baseline video age estimation results, to encourage and support further research on the topic of video-based age estimation, as the potential of this understudied field of research is clearly demonstrated.

Chapter 7

Conclusions

The estimation of chronological age based solely on visual cues is a challenging task. In fact, it is one of the most intricate problems in the face analysis research field. The main reason for this is the personalized and stochastic nature of the aging process, described in the first chapter. In this thesis, the problem of automatic age estimation is tackled using machine learning. Since machine learning is a very broad and complex research field, we dedicated the second chapter to introducing the most relevant theoretical foundations, related to learning algorithms, estimation methods, and the deep learning paradigm.

The sophistication and performance of modern-day machine learning models are truly impressive. However, it is not a common practice to estimate age from images or videos directly, via a single model. Multiple models, each serving a specific purpose, such as face detection, face alignment, and age estimation, are commonly organized in a processing pipeline. The general face analysis framework, with a special focus on the age estimation framework, is outlined in the third chapter. The abundance and variety of proposed algorithms and methods are evident from the overview of related work presented in the fourth chapter. To systematically review this fruitful field, we proposed a taxonomy that organizes the age estimation system design choices according to the type of input data, feature representation, and estimation algorithm. The early work was mostly focused on the design of novel feature representations, while in the era of deep learning the emphasis shifted towards the design of estimation algorithms. However, multiple studies indicate that the main contributor to the development of robust and precise age estimation systems is the training data.

Web scraping is a frequently utilized source for large quantities of diverse but unreliable data. The fifth chapter introduces an unsupervised biometric data filtering method capable of automatically reducing the number of erroneous samples in facial datasets. The experimental evaluation showed that the proposed method improved the performance significantly, even when more than half of the data was declared erroneous and discarded from training. As the lack of large and reliable public datasets is a common problem in the field, we further con-

tributed by introducing additional biometric filtering strategies and designing a newly derived public dataset, named Biometrically Filtered Famous Figure Dataset (B3FD). We demonstrated experimentally that it outperforms various state-of-the-art public datasets, while also showing that data quality is more important than data quantity, and that the choice of data contributes more than the choice of the age estimation method.

The potential of videos in the age estimation field is explored in the sixth chapter. While the benefits of using videos instead of static images were repeatedly discussed and experimentally demonstrated, the video-based methods are still far less studied. Once again, a big obstacle to the advancement of the video-based research field is the lack of appropriate data. We address the validation and comparability issues by designing a public video-based benchmark protocol, based on a recent in-the-wild evaluation dataset comprising a large number of videos and highly reliable labels. The second big obstacle, regarding the lack of training data, hinders the development of video-based methods that are able to utilize spatiotemporal dynamic information. We tackled this issue using a strategy based on semi-supervised learning and pseudo-labeling. The designed video-specific method, based on a temporal convolutional network, was shown to outperform its image-based counterpart, further corroborating the premise of video usefulness.

The principal contributions of this thesis are focused on data. The accuracy and robustness of machine learning models rely heavily on the quantity and quality of used training data. The validity and verifiability of research findings directly depend on the availability and quality of public benchmarks. The large performance disparities in the cross-dataset evaluation settings indicate the persistence of data-related issues in public datasets. In the era of advanced machine learning models, learning algorithms, and computational capabilities, the key to developing age estimation systems, that are truly robust in real-world conditions and to various intrinsic and extrinsic factors affecting facial appearance, is most likely to lie in the data itself.

Bibliography

[1]Liu, K.-H., Liu, T.-J., "A structure-based human facial age estimation framework under a constrained condition", IEEE Transactions on Image Processing, Vol. 28, No. 10, 2019, pp. 5187–5200.

[2]Agbo-Ajala, O., Viriri, S., "Deep learning approach for facial age classification: a survey of the state-of-the-art", Artificial Intelligence Review, Vol. 54, 2021, pp. 179–213.

[3]Al-Shannaq, A. S., Elrefaei, L. A., "Comprehensive analysis of the literature for age estimation from facial images", IEEE Access, Vol. 7, 2019, pp. 93 229–93 249.

[4]Angulu, R., Tapamo, J. R., Adewumi, A. O., "Age estimation via face images: a survey", EURASIP Journal on Image and Video Processing, Vol. 2018, No. 1, 2018, pp. 1–35.

[5]Farkas, J. P., Pessa, J. E., Hubbard, B., Rohrich, R. J., "The science and theory behind facial aging", Plastic and Reconstructive Surgery Global Open, Vol. 1, No. 1, 2013.

[6]Mohamed, E. G., Redondo, R. P. D., Koura, A., EL-Mofty, M. S., Kayed, M., "Dental age estimation using deep learning: A comparative survey", Computation, Vol. 11, No. 2, 2023, str. 18.

[7]Gnanasivam, P., Muttan, D. S., "Estimation of age through fingerprints using wavelet transform and singular value decomposition", International Journal of Biometrics and Bioinformatics (IJBB), Vol. 6, No. 2, 2012, pp. 58–67.

[8]Sharma, A., Rai, A., "An improved dcnn-based classification and automatic age estimation from multi-factorial mri data", in Advances in Computer, Communication and Computational Sciences: Proceedings of IC4S 2019. Springer, 2021, pp. 483–495.

[9]Štern, D., Payer, C., Urschler, M., "Automated age estimation from mri volumes of the hand", Medical image analysis, Vol. 58, 2019, str. 101538.

[10]Machado, C. E. P., Flores, M. R. P., Lima, L. N. C., Tinoco, R. L. R., Franco, A., Bezerra, A. C. B., Evison, M. P., Guimarães, M. A., "A new approach for the analysis of facial growth and age estimation: Iris ratio", PloS one, Vol. 12, No. 7, 2017.

[11]Harnsberger, J. D., Shrivastav, R., Brown Jr, W. S., Rothman, H., Hollien, H., "Speaking rate and fundamental frequency as speech cues to perceived age", Journal of voice, Vol. 22, No. 1, 2008, pp. 58–69.

[12]Lu, J., Tan, Y.-P., "Gait-based human age estimation", IEEE Transactions on Information Forensics and Security, Vol. 5, No. 4, 2010, pp. 761–770.

[13]Lanitis, A., "Age estimation based on head movements: A feasibility study", in 2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP). IEEE, 2010, pp. 1–6.

[14]Wang, M., Deng, W., "Deep face recognition: A survey", Neurocomputing, Vol. 429, 2021, pp. 215–244.

[15]Li, S., Deng, W., "Deep facial expression recognition: A survey", IEEE transactions on affective computing, Vol. 13, No. 3, 2020, pp. 1195–1215.

[16]Ghosh, S., Dhall, A., Hayat, M., Knibbe, J., Ji, Q., "Automatic gaze analysis: A survey of deep learning based approaches", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 46, No. 1, 2023, pp. 61–84.

[17]Ng, C.-B., Tay, Y.-H., Goi, B.-M., "A review of facial gender recognition", Pattern Analysis and Applications, Vol. 18, No. 4, 2015, pp. 739–755.

[18]Lu, X., Jain, A. K. *et al.*, "Ethnicity identification from face images", in Proceedings of SPIE, Vol. 5404. Citeseer, 2004, pp. 114–123.

[19]Hassan, B., Izquierdo, E., Piatrik, T., "Soft biometrics: a survey: Benchmark analysis, open challenges and recommendations", Multimedia Tools and Applications, 2021, pp. 1–44.

[20]Carletti, V., Greco, A., Percannella, G., Vento, M., "Age from faces in the deep learning revolution", IEEE transactions on pattern analysis and machine intelligence, Vol. 42, No. 9, 2019, pp. 2113–2132.

[21]Fu, Y., Guo, G., Huang, T. S., "Age synthesis and estimation via faces: A survey", IEEE transactions on pattern analysis and machine intelligence, Vol. 32, No. 11, 2010, pp. 1955–1976.

[22]Badr, M. M., Sarhan, A. M., Elbasiony, R. M., "Facial age estimation using deep neural networks: a survey", in 2019 15th International Computer Engineering Conference (ICENCO). IEEE, 2019, pp. 183–191.

[23]Geng, X., Zhou, Z.-H., Zhang, Y., Li, G., Dai, H., “Learning from facial aging patterns for automatic age estimation”, in Proceedings of the 14th ACM international conference on Multimedia, 2006, pp. 307–316.

[24]Mualla, N., Houssein, E. H., Zayed, H. H., “Face age estimation approach based on deep learning and principle component analysis”, International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018.

[25]El Dib, M. Y. M. H., “Automatic facial age estimation”, Unpublished master’s thesis, 2011.

[26]Bruyer, R., Scailquin, J.-C., “Person recognition and ageing: the cognitive status of addresses—an empirical question”, International Journal of Psychology, Vol. 29, No. 3, 1994, pp. 351–366.

[27]Anda, F., Lillis, D., Kanta, A., Becker, B. A., Bou-Harb, E., Le-Khac, N.-A., Scanlon, M., “Improving borderline adulthood facial age estimation through ensemble learning”, in Proceedings of the 14th International Conference on Availability, Reliability and Security, 2019, pp. 1–8.

[28]Agbo-Ajala, O., Viriri, S., Oloko-Oba, M., Ekundayo, O., Heymann, R., “Apparent age prediction from faces: A survey of modern approaches”, Frontiers in big Data, Vol. 5, 2022.

[29]Escalera, S., Fabian, J., Pardo, P., Baró, X., Gonzalez, J., Escalante, H. J., Misevic, D., Steiner, U., Guyon, I., “Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results”, in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 1–9.

[30]Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I., Rothe, R., “Apparent and real age estimation in still images with deep residual regressors on appa-real database”, in 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017, pp. 87–94.

[31]Padme, S. E., Desai, P., “Estimation of age from face images”, International Journal of Science and Research (IJSR), Vol. 4, No. 12, 2015.

[32]Rothe, R., Timofte, R., Van Gool, L., “Deep expectation of real and apparent age from a single image without facial landmarks”, International Journal of Computer Vision, Vol. 126, No. 2-4, 2018, pp. 144–157.

[33]Voelkle, M. C., Ebner, N. C., Lindenberger, U., Riediger, M., "Let me guess how old you are: effects of age, gender, and facial expression on perceptions of age.", Psychology and aging, Vol. 27, No. 2, 2012.

[34]Han, H., Otto, C., Jain, A. K., "Age estimation from face images: Human vs. machine performance", in Biometrics (ICB), 2013 International Conference on. IEEE, 2013, pp. 1–8.

[35]Lanitis, A., Taylor, C. J., Cootes, T. F., "Toward automatic simulation of aging effects on face images", IEEE Transactions on pattern Analysis and machine Intelligence, Vol. 24, No. 4, 2002, pp. 442–455.

[36]Geng, X., Yin, C., Zhou, Z.-H., "Facial age estimation by learning from label distributions", IEEE transactions on pattern analysis and machine intelligence, Vol. 35, No. 10, 2013, pp. 2401–2412.

[37]Guyuron, B., Rowe, D. J., Weinfeld, A. B., Eshraghi, Y., Fathi, A., Iamphongsai, S., "Factors contributing to the facial aging of identical twins", Plastic and reconstructive surgery, Vol. 123, No. 4, 2009, pp. 1321–1331.

[38]Sveikata, K., Balciuniene, I., Tutkuviene, J. *et al.*, "Factors influencing face aging. literature review", Stomatologija, Vol. 13, No. 4, 2011, pp. 113–116.

[39]Albert, A. M., Ricanek Jr, K., Patterson, E., "A review of the literature on the aging adult skull and face: Implications for forensic science research and applications", Forensic science international, Vol. 172, No. 1, 2007, pp. 1–9.

[40]Zimbler, M. S., Kokoska, M. S., Thomas, J. R., "Anatomy and pathophysiology of facial aging", Facial plastic surgery clinics of North America, Vol. 9, No. 2, 2001, pp. 179–187.

[41]Ghrban, Z. S., EL Abbadi, N. K., "Gender and age estimation from human faces based on deep learning techniques: A review", International Journal of Computing and Digital Systems, Vol. 14, No. 1, 2023, pp. 1–1.

[42]ELKarazle, K., Raman, V., Then, P., "Facial age estimation using machine learning techniques: An overview", Big Data and Cognitive Computing, Vol. 6, No. 4, 2022.

[43]Othmani, A., Taleb, A. R., Abdelkawy, H., Hadid, A., "Age estimation from faces using deep learning: A comparative analysis", Computer Vision and Image Understanding, Vol. 196, 2020.

[44]Shaw Jr, R. B., Katzel, E. B., Koltz, P. F., Kahn, D. M., Girotto, J. A., Langstein, H. N., "Aging of the mandible and its aesthetic implications", Plastic and reconstructive surgery, Vol. 125, No. 1, 2010, pp. 332–342.

[45]Taister, M. A., Holliday, S. D., Borrmman, H., "Comments on facial aging in law enforcement investigation", Forensic science communications, Vol. 2, No. 2, 2000.

[46]Vierkötter, A., Schikowski, T., Ranft, U., Sugiri, D., Matsui, M., Krämer, U., Krutmann, J., "Airborne particle exposure and extrinsic skin aging", Journal of investigative dermatology, Vol. 130, No. 12, 2010, pp. 2719–2726.

[47]Chen, C., Dantcheva, A., Ross, A., "Impact of facial cosmetics on automatic gender and age estimation algorithms", in 2014 International Conference on Computer Vision Theory and Applications (VISAPP), Vol. 2. IEEE, 2014, pp. 182–190.

[48]Chauhan, N., Warner, J. P., Adamson, P. A., "Perceived age change after aesthetic facial surgical procedures: quantifying outcomes of aging face surgery", Archives of Facial Plastic Surgery, Vol. 14, No. 4, 2012, pp. 258–262.

[49]Swanson, E., "Objective assessment of change in apparent age after facial rejuvenation surgery", Journal of Plastic, Reconstructive & Aesthetic Surgery, Vol. 64, No. 9, 2011, pp. 1124–1131.

[50]Nam, S. H., Kim, Y. H., Truong, N. Q., Choi, J., Park, K. R., "Age estimation by super-resolution reconstruction based on adversarial networks", IEEE Access, Vol. 8, 2020, pp. 17 103–17 120.

[51]Mahesh, B., "Machine learning algorithms-a review", International Journal of Science and Research (IJSR).[Internet], Vol. 9, No. 1, 2020, pp. 381–386.

[52]Goodfellow, I., Bengio, Y., Courville, A., Deep learning. MIT press, 2016.

[53]Mitchell, T. M., "Machine learning", 1997.

[54]Bishop, C. M., Nasrabadi, N. M., Pattern recognition and machine learning. Springer, 2006, Vol. 4, No. 4.

[55]Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2016.

[56]Torres-García, A. A., Garcia, C. A. R., Villasenor-Pineda, L., Mendoza-Montoya, O., Biosignal Processing and Classification Using Computational Learning and Intelligence: Principles, Algorithms, and Applications. Academic Press, 2021.

[57]Wold, S., Esbensen, K., Geladi, P., “Principal component analysis”, Chemometrics and intelligent laboratory systems, Vol. 2, No. 1-3, 1987, pp. 37–52.

[58]Biemann, C., “Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems”, in Proceedings of the first workshop on graph based methods for natural language processing. Association for Computational Linguistics, 2006, pp. 73–80.

[59]Frey, B. J., Dueck, D., “Clustering by passing messages between data points”, science, Vol. 315, No. 5814, 2007, pp. 972–976.

[60]Cheng, Y., “Mean shift, mode seeking, and clustering”, IEEE transactions on pattern analysis and machine intelligence, Vol. 17, No. 8, 1995, pp. 790–799.

[61]Van Engelen, J. E., Hoos, H. H., “A survey on semi-supervised learning”, Machine learning, Vol. 109, No. 2, 2020, pp. 373–440.

[62]Triguero, I., García, S., Herrera, F., “Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study”, Knowledge and Information systems, Vol. 42, 2015, pp. 245–284.

[63]LeCun, Y., Bengio, Y., Hinton, G., “Deep learning”, nature, Vol. 521, No. 7553, 2015, pp. 436–444.

[64]Werbos, P., “Beyond regression: New tools for prediction and analysis in the behavioral sciences”, PhD thesis, Committee on Applied Mathematics, Harvard University, Cambridge, MA, 1974.

[65]Rumelhart, D. E., Hinton, G. E., Williams, R. J., “Learning representations by back-propagating errors”, nature, Vol. 323, No. 6088, 1986, pp. 533–536.

[66]Hinton, G. E., Osindero, S., Teh, Y.-W., “A fast learning algorithm for deep belief nets”, Neural computation, Vol. 18, No. 7, 2006, pp. 1527–1554.

[67]Nair, V., Hinton, G. E., “Rectified linear units improve restricted boltzmann machines”, in Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.

[68]Hochreiter, S., “The vanishing gradient problem during learning recurrent neural nets and problem solutions”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 6, No. 02, 1998, pp. 107–116.

[69]Bengio, Y., Simard, P., Frasconi, P., “Learning long-term dependencies with gradient descent is difficult”, IEEE transactions on neural networks, Vol. 5, No. 2, 1994, pp. 157–166.

[70]Altabeiri, R., Alsafasfeh, M., Alhasanat, M., “Image compression approach for improving deep learning applications”, International Journal of Electrical and Computer Engineering (IJECE), Vol. 13, No. 5, 2023, pp. 5607–5616.

[71]Sarker, I. H., “Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions”, SN Computer Science, Vol. 2, No. 6, 2021.

[72]Rudolph, A., Krois, J., Hartmann, K., “Statistics and geodata analysis using python (soga-py)”, Department of Earth Sciences, Freie Universitaet Berlin, 2023.

[73]LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D., “Backpropagation applied to handwritten zip code recognition”, Neural computation, Vol. 1, No. 4, 1989, pp. 541–551.

[74]Dong, S., Wang, P., Abbas, K., “A survey on deep learning and its applications”, Computer Science Review, Vol. 40, 2021.

[75]Blauch, N. M., Behrmann, M., Plaut, D. C., “Computational insights into human perceptual expertise for familiar and unfamiliar face recognition”, Cognition, Vol. 208, 2021.

[76]LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., “Gradient-based learning applied to document recognition”, Proceedings of the IEEE, Vol. 86, No. 11, 1998, pp. 2278–2324.

[77]Krizhevsky, A., Sutskever, I., Hinton, G. E., “Imagenet classification with deep convolutional neural networks”, Advances in neural information processing systems, Vol. 25, 2012.

[78]Simonyan, K., Zisserman, A., “Very deep convolutional networks for large-scale image recognition”, arXiv preprint arXiv:1409.1556, 2014.

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

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

[81]Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K., "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size", arXiv preprint arXiv:1602.07360, 2016.

[82]Chollet, F., "Xception: Deep learning with depthwise separable convolutions", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[83]Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., "Mobilenets: Efficient convolutional neural networks for mobile vision applications", arXiv preprint arXiv:1704.04861, 2017.

[84]Chen, S., Liu, Y., Gao, X., Han, Z., "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices", in Chinese Conference on Biometric Recognition. Springer, 2018, pp. 428–438.

[85]Lea, C., Vidal, R., Reiter, A., Hager, G. D., "Temporal convolutional networks: A unified approach to action segmentation", in Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. Springer, 2016, pp. 47–54.

[86]Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. *et al.*, "Imagenet large scale visual recognition challenge", International journal of computer vision, Vol. 115, 2015, pp. 211–252.

[87]Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., "Imagenet: A large-scale hierarchical image database", in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[88]Parkhi, O., Vedaldi, A., Zisserman, A., "Deep face recognition", in BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association, 2015.

[89]Rothe, R., Timofte, R., Van Gool, L., "Deep expectation of real and apparent age from a single image without facial landmarks", International Journal of Computer Vision, 2016, pp. 1–14.

[90]Pan, H., Han, H., Shan, S., Chen, X., "Mean-variance loss for deep age estimation from a face", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5285–5294.

[91]Gao, B.-B., Zhou, H.-Y., Wu, J., Geng, X., "Age estimation using expectation of label distribution learning.", in IJCAI, 2018, pp. 712–718.

[92]Lin, H., Li, L., "Ordinal regression by extended binary classifications", Advances in neural information processing systems, Vol. 19, 2007, pp. 865–872.

[93]He, K., Zhang, X., Ren, S., Sun, J., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[94]Sun, Y., Wang, X., Tang, X., "Deeply learned face representations are sparse, selective, and robust", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2892–2900.

[95]Ricanek, K., Tesafaye, T., "Morph: A longitudinal image database of normal adult age-progression", in 7th international conference on automatic face and gesture recognition (FGR06). IEEE, 2006, pp. 341–345.

[96]Gupta, S. K., Nain, N., "Single attribute and multi attribute facial gender and age estimation", Multimedia Tools and Applications, Vol. 82, No. 1, 2023, pp. 1289–1311.

[97]Guo, G., Wang, X., "A study on human age estimation under facial expression changes", in 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 2547–2553.

[98]Nguyen, D. T., Cho, S. R., Shin, K. Y., Bang, J. W., Park, K. R. *et al.*, "Comparative study of human age estimation with or without preclassification of gender and facial expression", The Scientific World Journal, Vol. 2014, 2014.

[99]Bešeni ć, K., Gogić, I., Pandžić, I. S., Matković, K., "Automatic image-based face analysis systems overview", Engineering power: bulletin of the Croatian Academy of Engineering, Vol. 13, No. 2., 2018, pp. 2–7.

[100]Viola, P., Jones, M. J., "Robust real-time face detection", International journal of computer vision, Vol. 57, No. 2, 2004, pp. 137–154.

[101]Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L., "Face detection without bells and whistles", in European Conference on Computer Vision. Springer, 2014, pp. 720–735.

[102]Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G., "A convolutional neural network cascade for face detection", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5325–5334.

[103]Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A. C., "Ssd: Single shot multibox detector", in European conference on computer vision. Springer, 2016, pp. 21–37.

[104]Jiang, H., Learned-Miller, E., "Face detection with the faster r-cnn", in 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017). IEEE, 2017, pp. 650–657.

[105]Zhang, J., Wu, X., Hoi, S. C., Zhu, J., "Feature agglomeration networks for single stage face detection", Neurocomputing, Vol. 380, 2020, pp. 180–189.

[106]Zhang, S., Chi, C., Lei, Z., Li, S. Z., "Refineface: Refinement neural network for high performance face detection", IEEE transactions on pattern analysis and machine intelligence, Vol. 43, No. 11, 2020, pp. 4008–4020.

[107]Zhang, C., Zhang, Z., "A survey of recent advances in face detection", 2010.

[108]Kumar, A., Kaur, A., Kumar, M., "Face detection techniques: a review", Artificial Intelligence Review, Vol. 52, 2019, pp. 927–948.

[109]Minaee, S., Luo, P., Lin, Z., Bowyer, K., "Going deeper into face detection: A survey", arXiv preprint arXiv:2103.14983, 2021.

[110]Antipov, G., Baccouche, M., Berrani, S.-A., Dugelay, J.-L., "Apparent age estimation from face images combining general and children-specialized deep learning models", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 96–104.

[111]Wu, Y., Ji, Q., "Facial landmark detection: A literature survey", International Journal of Computer Vision, Vol. 127, 2019, pp. 115–142.

[112]Gogi ć, I., Ahlberg, J., Pandžić, I. S., "Regression-based methods for face alignment: A survey", Signal Processing, Vol. 178, 2021.

[113]Khabarlak, K., Koriashkina, L., "Fast facial landmark detection and applications: A survey", arXiv preprint arXiv:2101.10808, 2021.

[114]Han, H., Otto, C., Liu, X., Jain, A. K., "Demographic estimation from face images: Human vs. machine performance", IEEE transactions on pattern analysis and machine intelligence, Vol. 37, No. 6, 2015, pp. 1148–1161.

[115]Hu, Z., Wen, Y., Wang, J., Wang, M., Hong, R., Yan, S., "Facial age estimation with age difference", IEEE Transactions on Image Processing, Vol. 26, No. 7, 2017, pp. 3087–3097.

[116]Han, H., Jain, A. K., "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), 2014.

[117]Bundy, A., Wallen, L., "Difference of gaussians", Catalogue of Artificial Intelligence Tools, 1984, pp. 30–30.

[118]Zuiderveld, K., "Contrast limited adaptive histogram equalization", Graphics gems, 1994, pp. 474–485.

[119]Ahonen, T., Hadid, A., Pietikainen, M., "Face description with local binary patterns: Application to face recognition", IEEE transactions on pattern analysis and machine intelligence, Vol. 28, No. 12, 2006, pp. 2037–2041.

[120]Meyers, E., Wolf, L., "Using biologically inspired features for face processing", International Journal of Computer Vision, Vol. 76, 2008, pp. 93–104.

[121]Dalal, N., Triggs, B., "Histograms of oriented gradients for human detection", in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Vol. 1. Ieee, 2005, pp. 886–893.

[122]Bay, H., Tuytelaars, T., Van Gool, L., "Surf: Speeded up robust features", in Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9. Springer, 2006, pp. 404–417.

[123]Balakrishnama, S., Ganapathiraju, A., "Linear discriminant analysis-a brief tutorial", Institute for Signal and information Processing, Vol. 18, No. 1998, 1998, pp. 1–8.

[124]Boser, B. E., Guyon, I. M., Vapnik, V. N., "A training algorithm for optimal margin classifiers", in Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.

[125]Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., "Support vector regression machines", Advances in neural information processing systems, Vol. 9, 1996.

[126]Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., Feuston, B. P., "Random forest: a classification and regression tool for compound classification and qsar modeling", Journal of chemical information and computer sciences, Vol. 43, No. 6, 2003, pp. 1947–1958.

[127]Peterson, L. E., "K-nearest neighbor", Scholarpedia, Vol. 4, No. 2, 2009.

[128]Freund, Y., Schapire, R. E. *et al.*, "Experiments with a new boosting algorithm", in icml, Vol. 96. Citeseer, 1996, pp. 148–156.

[129]Gallagher, A. C., Chen, T., "Understanding images of groups of people", in 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 256–263.

[130]Eidinger, E., Enbar, R., Hassner, T., "Age and gender estimation of unfiltered faces", IEEE Transactions on Information Forensics and Security, Vol. 9, No. 12, 2014, pp. 2170–2179.

[131]Kwon, Y. H. *et al.*, "Age classification from facial images", in Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on. IEEE, 1994, pp. 762–767.

[132]Golomb, B. A., Lawrence, D. T., Sejnowski, T. J., "Sexnet: A neural network identifies sex from human faces.", in NIPS, Vol. 1, 1990.

[133]Panis, G., Lanitis, A., "An overview of research activities in facial age estimation using the fg-net aging database", in European Conference on Computer Vision. Springer, 2014, pp. 737–750.

[134]Escalera, S., Torres Torres, M., Martinez, B., Baró, X., Jair Escalante, H., Guyon, I., Tzimiropoulos, G., Corneou, C., Oliu, M., Ali Bagheri, M. *et al.*, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 1–8.

[135]Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S., "Agedb: the first manually collected, in-the-wild age database", in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 51–59.

[136]Zhang, Y., Liu, L., Li, C. *et al.*, "Quantifying facial age by posterior of age comparisons", arXiv preprint arXiv:1708.09687, 2017.

[137]Zhang, K., Gao, C., Guo, L., Sun, M., Yuan, X., Han, T. X., Zhao, Z., Li, B., "Age group and gender estimation in the wild with deep ror architecture", IEEE Access, Vol. 5, 2017, pp. 22 492–22 503.

[138]Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G., "Ordinal regression with multiple output cnn for age estimation", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4920–4928.

[139]Chen, B.-C., Chen, C.-S., Hsu, W. H., "Cross-age reference coding for age-invariant face recognition and retrieval", in European Conference on Computer Vision. Springer, 2014, pp. 768–783.

[140]Rothe, R., Timofte, R., Van Gool, L., "Dex: Deep expectation of apparent age from a single image", in Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 10–15.

[141]Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., Brossard, E., "The megaface benchmark: 1 million faces for recognition at scale", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4873–4882.

[142]Ni, K., Pearce, R., Boakye, K., Van Essen, B., Borth, D., Chen, B., Wang, E., "Large-scale deep learning on the yfcc100m dataset", arXiv preprint arXiv:1502.03409, 2015.

[143]Dibeklio ğlu, H., Salah, A. A., Gevers, T., "Are you really smiling at me? spontaneous versus posed enjoyment smiles", in Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12. Springer, 2012, pp. 525–538.

[144]Dibeklio ğlu, H., Gevers, T., Salah, A. A., Valenti, R., "A smile can reveal your age: Enabling facial dynamics in age estimation", in Proceedings of the 20th ACM international conference on Multimedia, 2012, pp. 209–218.

[145]Dibeklio ğlu, H., Alnajar, F., Salah, A. A., Gevers, T., "Combining facial dynamics with appearance for age estimation", IEEE Transactions on Image Processing, Vol. 24, No. 6, 2015, pp. 1928–1943.

[146]Han, J., Wang, W., Karaoglu, S., Zeng, W., Gevers, T., "Pose invariant age estimation of face images in the wild", Computer Vision and Image Understanding, Vol. 202, 2021.

[147]Hazirbas, C., Bitton, J., Dolhansky, B., Pan, J., Gordo, A., Ferrer, C. C., "Towards measuring fairness in ai: the casual conversations dataset", IEEE Transactions on Biometrics, Behavior, and Identity Science, Vol. 4, No. 3, 2021, pp. 324–332.

[148]Fitzpatrick, T. B., "Soleil et peau", J. Med. Esthet., Vol. 2, 1975, pp. 33–34.

[149]Porgali, B., Albiero, V., Ryda, J., Ferrer, C. C., Hazirbas, C., "The casual conversations v2 dataset", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10–17.

[150]Monk, J., Ellis P., "Skin Tone Stratification among Black Americans, 2001–2003", Social Forces, Vol. 92, No. 4, 03 2014, pp. 1313-1337.

[151]Zhou, S. K., Georgescu, B., Zhou, X. S., Comaniciu, D., "Image based regression using boosting method", in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Vol. 1. IEEE, 2005, pp. 541–548.

[152]Hastie, T., Tibshirani, R., Friedman, J. H., Friedman, J. H., The elements of statistical learning: data mining, inference, and prediction. Springer, 2009, Vol. 2.

[153]Yan, S., Wang, H., Tang, X., Huang, T. S., "Learning auto-structured regressor from uncertain nonnegative labels", in 2007 IEEE 11th international conference on computer vision. IEEE, 2007, pp. 1–8.

[154]Lanitis, A., Draganova, C., Christodoulou, C., "Comparing different classifiers for automatic age estimation", IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 34, No. 1, 2004, pp. 621–628.

[155]Suo, J., Wu, T., Zhu, S., Shan, S., Chen, X., Gao, W., "Design sparse features for age estimation using hierarchical face model", in 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, 2008, pp. 1–6.

[156]Guo, G., Fu, Y., Huang, T. S., Dyer, C. R., "Locally adjusted robust regression for human age estimation", in 2008 IEEE Workshop on Applications of Computer Vision. IEEE, 2008, pp. 1–6.

[157]Fu, Y., Xu, Y., Huang, T. S., "Estimating human age by manifold analysis of face pictures and regression on aging features", in 2007 IEEE International Conference on Multimedia and Expo. IEEE, 2007, pp. 1383–1386.

[158]Fu, Y., Huang, T. S., "Human age estimation with regression on discriminative aging manifold", IEEE Transactions on Multimedia, Vol. 10, No. 4, 2008, pp. 578–584.

[159]Weisberg, S., Applied linear regression. John Wiley & Sons, 2005, Vol. 528.

[160]Guo, G., Mu, G., "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression", in CVPR 2011. IEEE, 2011, pp. 657–664.

[161]Chao, W.-L., Liu, J.-Z., Ding, J.-J., "Facial age estimation based on label-sensitive learning and age-oriented regression", Pattern Recognition, Vol. 46, No. 3, 2013, pp. 628–641.

[162]Cai, L., Huang, L., Liu, C., "Age estimation based on improved discriminative gaussian process latent variable model", Multimedia Tools and Applications, Vol. 75, 2016, pp. 11 977–11 994.

[163]Guo, G., Fu, Y., Dyer, C. R., Huang, T. S., "Image-based human age estimation by manifold learning and locally adjusted robust regression", IEEE Transactions on Image Processing, Vol. 17, No. 7, 2008, pp. 1178–1188.

[164]Guo, G., Fu, Y., Dyer, C. R., Huang, T. S., “A probabilistic fusion approach to human age prediction”, in 2008 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE, 2008, pp. 1–6.

[165]Guo, G., Mu, G., Fu, Y., Huang, T. S., “Human age estimation using bio-inspired features”, in 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009, pp. 112–119.

[166]Huerta, I., Fernández, C., Segura, C., Hernando, J., Prati, A., “A deep analysis on age estimation”, Pattern Recognition Letters, Vol. 68, 2015, pp. 239–249.

[167]Yi, D., Lei, Z., Li, S. Z., “Age estimation by multi-scale convolutional network”, in Asian conference on computer vision. Springer, 2014, pp. 144–158.

[168]Dornaika, F., Bekhouche, S. E., Arganda-Carreras, I., “Robust regression with deep cnns for facial age estimation: An empirical study”, Expert Systems with Applications, Vol. 141, 2020.

[169]Rothe, R., Timofte, R., Van Gool, L., “Some like it hot - visual guidance for preference prediction”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[170]Zhu, Y., Li, Y., Mu, G., Guo, G., “A study on apparent age estimation”, in Computer Vision Workshop (ICCVW), 2015 IEEE International Conference on. IEEE, 2015, pp. 267–273.

[171]Breiman, L., “Random forests”, Machine learning, Vol. 45, 2001, pp. 5–32.

[172]Kwon, Y. H., da Vitoria Lobo, N., “Age classification from facial images”, Computer vision and image understanding, Vol. 74, No. 1, 1999, pp. 1–21.

[173]Dehshibi, M. M., Bastanfard, A., “A new algorithm for age recognition from facial images”, Signal Processing, Vol. 90, No. 8, 2010, pp. 2431–2444.

[174]Hajizadeh, M. A., Ebrahimnezhad, H., “Classification of age groups from facial image using histograms of oriented gradients”, in 2011 7th Iranian conference on machine vision and image processing. IEEE, 2011, pp. 1–5.

[175]Wang, J.-G., Sung, E., Yau, W.-Y., “Active learning for solving the incomplete data problem in facial age classification by the furthest nearest-neighbor criterion”, IEEE transactions on image processing, Vol. 20, No. 7, 2011, pp. 2049–2062.

[176]Han, H., Otto, C., Liu, X., Jain, A. K., “Demographic estimation from face images: Human vs. machine performance”, IEEE transactions on pattern analysis and machine intelligence, Vol. 37, No. 6, 2014, pp. 1148–1161.

[177]Sawant, M. M., Bhurchandi, K., “Hierarchical facial age estimation using gaussian process regression”, IEEE Access, Vol. 7, 2019, pp. 9142–9152.

[178]Tokola, R., Bolme, D., Boehnen, C., Barstow, D., Ricanek, K., “Discriminating projections for estimating face age in wild images”, in IEEE International Joint Conference on Biometrics. IEEE, 2014, pp. 1–8.

[179]Guo, G., Mu, G., Fu, Y., Dyer, C., Huang, T., “A study on automatic age estimation using a large database”, in 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009, pp. 1986–1991.

[180]Liu, K.-H., Yan, S., Kuo, C.-C. J., “Age estimation via grouping and decision fusion”, IEEE Transactions on Information Forensics and Security, Vol. 10, No. 11, 2015, pp. 2408–2423.

[181]Ozbulak, G., Aytar, Y., Ekenel, H. K., “How transferable are cnn-based features for age and gender classification?”, in 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2016, pp. 1–6.

[182]Uricár, M., Timofte, R., Rothe, R., Matas, J., Van Gool, L., “Structured output svm prediction of apparent age, gender and smile from deep features”, in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 25–33.

[183]Qawaqneh, Z., Mallouh, A. A., Barkana, B. D., “Deep convolutional neural network for age estimation based on vgg-face model”, arXiv preprint arXiv:1709.01664, 2017.

[184]Hebda, B., Kryjak, T., “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), 2016, pp. 787-790.

[185]Levi, G., Hassner, T., “Age and gender classification using convolutional neural networks”, in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2015, pp. 34–42.

[186]Antipov, G., Baccouche, M., Berrani, S.-A., Dugelay, J.-L., “Effective training of convolutional neural networks for face-based gender and age prediction”, Pattern Recognition, Vol. 72, 2017, pp. 15–26.

[187]Della Pietra, S., Della Pietra, V., Lafferty, J., "Inducing features of random fields", IEEE transactions on pattern analysis and machine intelligence, Vol. 19, No. 4, 1997, pp. 380–393.

[188]Zeng, X.-Q., Xiang, R., Zou, H.-X., "Partial least squares regression based facial age estimation", in 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Vol. 1, 2017, pp. 416-421.

[189]Yang, X., Gao, B.-B., Xing, C., Huo, Z.-W., Wei, X.-S., Zhou, Y., Wu, J., Geng, X., "Deep label distribution learning for apparent age estimation", in Proceedings of the IEEE international conference on computer vision workshops, 2015, pp. 102–108.

[190]Gao, B.-B., Xing, C., Xie, C.-W., Wu, J., Geng, X., "Deep label distribution learning with label ambiguity", IEEE Transactions on Image Processing, Vol. 26, No. 6, 2017, pp. 2825–2838.

[191]Li, P., Hu, Y., Wu, X., He, R., Sun, Z., "Deep label refinement for age estimation", Pattern Recognition, Vol. 100, 2020.

[192]He, Z., Li, X., Zhang, Z., Wu, F., Geng, X., Zhang, Y., Yang, M.-H., Zhuang, Y., "Data-dependent label distribution learning for age estimation", IEEE Transactions on Image processing, Vol. 26, No. 8, 2017, pp. 3846–3858.

[193]Gao, B.-B., "Jointly learning distribution and expectation in a unified framework for facial age and attractiveness estimation", Neural Computing and Applications, Vol. 35, No. 21, 2023, pp. 15 583–15 599.

[194]Chang, K.-Y., Chen, C.-S., Hung, Y.-P., "A ranking approach for human ages estimation based on face images", in 2010 20th International Conference on Pattern Recognition. IEEE, 2010, pp. 3396–3399.

[195]Chang, K.-Y., Chen, C.-S., Hung, Y.-P., "Ordinal hyperplanes ranker with cost sensitivities for age estimation", in CVPR 2011. IEEE, 2011, pp. 585–592.

[196]Li, L., Lin, H.-T., "Ordinal regression by extended binary classification", Advances in neural information processing systems, Vol. 19, 2006.

[197]Chang, K.-Y., Chen, C.-S., "A learning framework for age rank estimation based on face images with scattering transform", IEEE Transactions on Image Processing, Vol. 24, No. 3, 2015, pp. 785–798.

[198]Li, C., Liu, Q., Dong, W., Zhu, X., Liu, J., Lu, H., "Human age estimation based on locality and ordinal information", IEEE transactions on cybernetics, Vol. 45, No. 11, 2014, pp. 2522–2534.

[199]Yang, H.-F., Lin, B.-Y., Chang, K.-Y., Chen, C.-S. *et al.*, "Automatic age estimation from face images via deep ranking", networks, Vol. 35, No. 8, 2013, pp. 1872–1886.

[200]Zeng, X., Huang, J., Ding, C., "Soft-ranking label encoding for robust facial age estimation", IEEE Access, Vol. 8, 2020, pp. 134 209–134 218.

[201]Chen, S., Zhang, C., Dong, M., "Deep age estimation: From classification to ranking", IEEE Transactions on Multimedia, Vol. 20, No. 8, 2017, pp. 2209–2222.

[202]Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., Han, H., Shan, S., Chen, X., "Agenet: Deeply learned regressor and classifier for robust apparent age estimation", in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 16–24.

[203]Diaz, R., Marathe, A., "Soft labels for ordinal regression", in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4738–4747.

[204]Li, Q., Wang, J., Yao, Z., Li, Y., Yang, P., Yan, J., Wang, C., Pu, S., "Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20 513–20 522.

[205]Farkas, L. G., "Anthropometry of the head and face", New York: Raven Press, 1994.

[206]Horng, W.-B., Lee, C.-P., Chen, C.-W. *et al.*, "Classification of age groups based on facial features", Journal of applied science and engineering, Vol. 4, No. 3, 2001, pp. 183–192.

[207]Ramanathan, N., Chellappa, R., "Modeling age progression in young faces", in 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), Vol. 1. IEEE, 2006, pp. 387–394.

[208]Turaga, P., Biswas, S., Chellappa, R., "The role of geometry in age estimation", in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010, pp. 946–949.

[209]Thukral, P., Mitra, K., Chellappa, R., "A hierarchical approach for human age estimation", in 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2012, pp. 1529–1532.

[210]Cootes, T. F., Edwards, G. J., Taylor, C. J., “Active appearance models”, IEEE Transactions on pattern analysis and machine intelligence, Vol. 23, No. 6, 2001, pp. 681–685.

[211]Yan, S., Wang, H., Huang, T. S., Yang, Q., Tang, X., “Ranking with uncertain labels”, in 2007 IEEE international conference on multimedia and expo. IEEE, 2007, pp. 96–99.

[212]Chen, K., Gong, S., Xiang, T., Change Loy, C., “Cumulative attribute space for age and crowd density estimation”, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2467–2474.

[213]Feng, S., Lang, C., Feng, J., Wang, T., Luo, J., “Human facial age estimation by cost-sensitive label ranking and trace norm regularization”, IEEE Transactions on Multimedia, Vol. 19, No. 1, 2016, pp. 136–148.

[214]Geng, X., Zhou, Z.-H., Smith-Miles, K., “Automatic age estimation based on facial aging patterns”, IEEE Transactions on pattern analysis and machine intelligence, Vol. 29, No. 12, 2007, pp. 2234–2240.

[215]Scherbaum, K., Sunkel, M., Seidel, H.-P., Blanz, V., “Prediction of individual non-linear aging trajectories of faces”, in Computer Graphics Forum, Vol. 26, No. 3. Wiley Online Library, 2007, pp. 285–294.

[216]Yan, S., Wang, H., Fu, Y., Yan, J., Tang, X., Huang, T. S., “Synchronized submanifold embedding for person-independent pose estimation and beyond”, IEEE transactions on image processing, Vol. 18, No. 1, 2008, pp. 202–210.

[217]Yang, Z., Ai, H., “Demographic classification with local binary patterns”, Advances in Biometrics, 2007, pp. 464–473.

[218]Gunay, A., Nabiyev, V. V., “Automatic age classification with lbp”, in 2008 23rd international symposium on computer and information sciences. IEEE, 2008, pp. 1–4.

[219]Unnikrishnan, A., Ajesh, F., Kizhakkethottam, J. J., “Texture-based estimation of age and gender from wild conditions”, Procedia Technology, Vol. 24, 2016, pp. 1349–1357.

[220]Gao, F., Ai, H., “Face age classification on consumer images with gabor feature and fuzzy lda method”, in Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3. Springer, 2009, pp. 132–141.

[221]Riesenhuber, M., Poggio, T., “Hierarchical models of object recognition in cortex”, Nature neuroscience, Vol. 2, No. 11, 1999, pp. 1019–1025.

[222]El Dib, M. Y., El-Saban, M., “Human age estimation using enhanced bio-inspired features (ebif)”, in Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE, 2010, pp. 1589–1592.

[223]Hayashi, J., Yasumoto, M., Ito, H., Koshimizu, H., “Method for estimating and modeling age and gender using facial image processing”, in Proceedings Seventh International Conference on Virtual Systems and Multimedia. IEEE, 2001, pp. 439–448.

[224]Belver, C., Arganda-Carreras, I., Dornaika, F., “Comparative study of human age estimation based on hand-crafted and deep face features”, in Video Analytics. Face and Facial Expression Recognition and Audience Measurement: Third International Workshop, VAAM 2016, and Second International Workshop, FFER 2016, Cancun, Mexico, December 4, 2016, Revised Selected Papers 2. Springer, 2017, pp. 98–112.

[225]Bianco, S., Cadene, R., Celona, L., Napoletano, P., “Benchmark analysis of representative deep neural network architectures”, IEEE access, Vol. 6, 2018, pp. 64 270–64 277.

[226]Dong, Y., Liu, Y., Lian, S., “Automatic age estimation based on deep learning algorithm”, Neurocomputing, Vol. 187, 2016, pp. 4–10.

[227]Xing, J., Li, K., Hu, W., Yuan, C., Ling, H., “Diagnosing deep learning models for high accuracy age estimation from a single image”, Pattern Recognition, Vol. 66, 2017, pp. 106–116.

[228]Li, K., Xing, J., Hu, W., Maybank, S. J., “D2c: Deep cumulatively and comparatively learning for human age estimation”, Pattern Recognition, Vol. 66, 2017, pp. 95–105.

[229]Han, H., Jain, A. K., Wang, F., Shan, S., Chen, X., “Heterogeneous face attribute estimation: A deep multi-task learning approach”, IEEE transactions on pattern analysis and machine intelligence, Vol. 40, No. 11, 2017, pp. 2597–2609.

[230]Liu, H., Lu, J., Feng, J., Zhou, J., “Ordinal deep learning for facial age estimation”, IEEE transactions on circuits and systems for video technology, Vol. 29, No. 2, 2017, pp. 486–501.

[231]Hu, Z., Wen, Y., Wang, J., Wang, M., Hong, R., Yan, S., “Facial age estimation with age difference”, IEEE Transactions on Image Processing, Vol. 26, No. 7, 2016, pp. 3087–3097.

[232]Berg, A., Oskarsson, M., O’Connor, M., “Deep ordinal regression with label diversity”, in 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 2740–2747.

[233]Cao, W., Mirjalili, V., Raschka, S., "Rank consistent ordinal regression for neural networks with application to age estimation", Pattern Recognition Letters, Vol. 140, 2020, pp. 325–331.

[234]Liu, H., Lu, J., Feng, J., Zhou, J., "Label-sensitive deep metric learning for facial age estimation", IEEE Transactions on Information Forensics and Security, Vol. 13, No. 2, 2017, pp. 292–305.

[235]Doždor, Z., Hrka ć, T., Brkić, K., Kalafatić, Z., "Facial age estimation models for embedded systems: A comparative study", IEEE Access, Vol. 11, 2023, pp. 14 282–14 292.

[236]Yang, T.-Y., Huang, Y.-H., Lin, Y.-Y., Hsiu, P.-C., Chuang, Y.-Y., "Ssr-net: A compact soft stagewise regression network for age estimation.", in IJCAI, Vol. 5, No. 6, 2018.

[237]Bruna, J., Mallat, S., "Invariant scattering convolution networks", IEEE transactions on pattern analysis and machine intelligence, Vol. 35, No. 8, 2013, pp. 1872–1886.

[238]Zhang, C., Liu, S., Xu, X., Zhu, C., "C3ae: Exploring the limits of compact model for age estimation", in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12 587–12 596.

[239]Zhang, K., Liu, N., Yuan, X., Guo, X., Gao, C., Zhao, Z., Ma, Z., "Fine-grained age estimation in the wild with attention lstm networks", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 30, No. 9, 2019, pp. 3140–3152.

[240]Paplham, J., Franc, V., "A call to reflect on evaluation practices for age estimation: Comparative analysis of the state-of-the-art and a unified benchmark", arXiv preprint arXiv:2307.04570, 2024.

[241]Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F., "General facial representation learning in a visual-linguistic manner", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18 697–18 709.

[242]Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale", arXiv preprint arXiv:2010.11929, 2020.

[243]Yang, M., Zhu, S., Lv, F., Yu, K., "Correspondence driven adaptation for human profile recognition", in CVPR 2011. IEEE, 2011, pp. 505–512.

[244]Wan, J., Tan, Z., Lei, Z., Guo, G., Li, S. Z., "Auxiliary demographic information assisted age estimation with cascaded structure", IEEE transactions on cybernetics, Vol. 48, No. 9, 2018, pp. 2531–2541.

[245]Hou, L., Yu, C.-P., Samaras, D., "Squared earth mover's distance-based loss for training deep neural networks", arXiv preprint arXiv:1611.05916, 2016.

[246]Kuang, Z., Huang, C., Zhang, W., "Deeply learned rich coding for cross-dataset facial age estimation", in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 96–101.

[247]Cao, Q., Shen, L., Xie, W., Parkhi, O. M., Zisserman, A., "Vggface2: A dataset for recognising faces across pose and age", in Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on. IEEE, 2018, pp. 67–74.

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

[249]Huo, Z., Yang, X., Xing, C., Zhou, Y., Hou, P., Lv, J., Geng, X., "Deep age distribution learning for apparent age estimation", in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 17–24.

[250]Hou, L., Samaras, D., Kurc, T., Gao, Y., Saltz, J., "Convnets with smooth adaptive activation functions for regression", in Artificial Intelligence and Statistics. PMLR, 2017, pp. 430–439.

[251]Hadid, A., "Analyzing facial behavioral features from videos", in Human Behavior Understanding: Second International Workshop, HBU 2011, Amsterdam, The Netherlands, November 16, 2011. Proceedings 2. Springer, 2011, pp. 52–61.

[252]Ojala, T., Pietikäinen, M., Harwood, D., "A comparative study of texture measures with classification based on featured distributions", Pattern recognition, Vol. 29, No. 1, 1996, pp. 51–59.

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

[254]Zhao, G., Pietikainen, M., "Dynamic texture recognition using local binary patterns with an application to facial expressions", IEEE transactions on pattern analysis and machine intelligence, Vol. 29, No. 6, 2007, pp. 915–928.

[255]Alnajar, F., Shan, C., Gevers, T., Geusebroek, J.-M., "Learning-based encoding with soft assignment for age estimation under unconstrained imaging conditions", Image and Vision Computing, Vol. 30, No. 12, 2012, pp. 946–953.

[256]Pei, W., Dibeklio ğlu, H., Baltrušaitis, T., Tax, D. M., "Attended end-to-end architecture for age estimation from facial expression videos", IEEE Transactions on Image Processing, Vol. 29, 2019, pp. 1972–1984.

[257]Ji, Z., Lang, C., Li, K., Xing, J., "Deep age estimation model stabilization from images to videos", in 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018, pp. 1420–1425.

[258]Zhang, B., Bao, Y., "Age estimation of faces in videos using head pose estimation and convolutional neural networks", Sensors, Vol. 22, No. 11, 2022.

[259]Ruiz, N., Chong, E., Rehg, J. M., "Fine-grained head pose estimation without keypoints", in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 2074–2083.

[260]Shen, W., Guo, Y., Wang, Y., Zhao, K., Wang, B., Yuille, A. L., "Deep regression forests for age estimation", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2304–2313.

[261]Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S. Z., "Face alignment across large poses: A 3d solution", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 146–155.

[262]Chen, B.-C., Chen, C.-S., Hsu, W. H., "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset", IEEE Transactions on Multimedia, Vol. 17, No. 6, 2015, pp. 804–815.

[263]Panis, G., Lanitis, A., Tsapatsoulis, N., Cootes, T. F., "Overview of research on facial ageing using the fg-net ageing database", Iet Biometrics, Vol. 5, No. 2, 2016, pp. 37–46.

[264]Jia, S., Cristianini, N., "Learning to classify gender from four million images", Pattern recognition letters, Vol. 58, 2015, pp. 35–41.

[265]Ni, B., Song, Z., Yan, S., "Web image mining towards universal age estimator", in Proceedings of the 17th ACM international conference on Multimedia. ACM, 2009, pp. 85–94.

[266]Bešeni ć, K., Ahlberg, J., Pandžić, I. S., "Unsupervised facial biometric data filtering for age and gender estimation", in Proceedings of the 14th International Joint Conference on

Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, 2019, pp. 209-217.

[267]Bešeni ć, K., Ahlberg, J., Pandžić, I. S., "Picking out the bad apples: unsupervised biometric data filtering for refined age estimation", The Visual Computer, Vol. 39, No. 1, Jan 2023, pp. 219-237.

[268]Huang, G. B., Mattar, M., Berg, T., Learned-Miller, E., "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments", in Workshop on faces in'Real-Life'Images: detection, alignment, and recognition, 2008.

[269]Ng, H.-W., Winkler, S., "A data-driven approach to cleaning large face datasets", in Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, 2014, pp. 343–347.

[270]Bulat, A., Tzimiropoulos, G., "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)", in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1021–1030.

[271]Redmon, J., Farhadi, A., "Yolo9000: better, faster, stronger", in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271.

[272]Zeiler, M. D., "Adadelta: an adaptive learning rate method", arXiv preprint arXiv:1212.5701, 2012.

[273]Bassili, J. N., "Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face.", Journal of personality and social psychology, Vol. 37, No. 11, 1979.

[274]Hill, H., Johnston, A., "Categorizing sex and identity from the biological motion of faces", Current biology, Vol. 11, No. 11, 2001, pp. 880–885.

[275]Knight, B., Johnston, A., "The role of movement in face recognition", Visual cognition, Vol. 4, No. 3, 1997, pp. 265–273.

[276]O'Toole, A. J., Roark, D. A., Abdi, H., "Recognizing moving faces: A psychological and neural synthesis", Trends in cognitive sciences, Vol. 6, No. 6, 2002, pp. 261–266.

[277]Bešeni ć., K., Pandžić., I., Ahlberg., J., "Let me take a better look: Towards video-based age estimation", in Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods - ICPRAM, INSTICC. SciTePress, 2024, pp. 57-69.

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

[279]Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., Zuiderveld, K., "Adaptive histogram equalization and its variations", Computer vision, graphics, and image processing, Vol. 39, No. 3, 1987, pp. 355–368.

[280]Lee, J.-H., Chan, Y.-M., Chen, T.-Y., Chen, C.-S., "Joint estimation of age and gender from unconstrained face images using lightweight multi-task cnn for mobile applications", in 2018 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, 2018, pp. 162–165.

[281]Serengil, S. I., Ozpinar, A., "Lightface: A hybrid deep face recognition framework", in 2020 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, 2020, pp. 1–5.

[282]Gao, J., Yang, Z., Nevatia, R., "Red: Reinforced encoder-decoder networks for action anticipation", arXiv preprint arXiv:1707.04818, 2017.

[283]Xu, M., Gao, M., Chen, Y.-T., Davis, L. S., Crandall, D. J., "Temporal recurrent networks for online action detection", in IEEE International Conference on Computer Vision (ICCV), 2019.

[284]Eun, H., Moon, J., Park, J., Jung, C., Kim, C., "Temporal filtering networks for online action detection", Pattern Recognition, Vol. 111, 2021.

[285]Xu, M., Xiong, Y., Chen, H., Li, X., Xia, W., Tu, Z., Soatto, S., "Long short-term transformer for online action detection", Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 1086–1099.

[286]Chen, J., Mittal, G., Yu, Y., Kong, Y., Chen, M., "Gatehub: Gated history unit with background suppression for online action detection", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19 925–19 934.

[287]Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., Goodfellow, I., "Realistic evaluation of deep semi-supervised learning algorithms", Advances in neural information processing systems, Vol. 31, 2018.

[288]Hu, Z., Yang, Z., Hu, X., Nevatia, R., "Simple: Similar pseudo label exploitation for semi-supervised classification", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15 099–15 108.

[289]Maze, B., Adams, J., Duncan, J. A., Kalka, N., Miller, T., Otto, C., Jain, A. K., Niggel, W. T., Anderson, J., Cheney, J. *et al.*, "Iarpa janus benchmark-c: Face dataset and protocol", in 2018 international conference on biometrics (ICB). IEEE, 2018, pp. 158–165.

[290]Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C. C., "CelebV-HQ: A large-scale video facial attributes dataset", in ECCV, 2022.

[291]De Geest, R., Gavves, E., Ghodrati, A., Li, Z., Snoek, C., Tuytelaars, T., "Online action detection", in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer, 2016, pp. 269–284.

[292]De Geest, R., Tuytelaars, T., "Modeling temporal structure with lstm for online action detection", in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 1549–1557.

[293]Kim, Y. H., Nam, S., Kim, S. J., "Temporally smooth online action detection using cycle-consistent future anticipation", Pattern Recognition, Vol. 116, 2021.

[294]Wang, W., Peng, X., Qiao, Y., Cheng, J., "An empirical study on temporal modeling for online action detection", Complex & Intelligent Systems, Vol. 8, No. 2, 2022, pp. 1803–1817.

[295]Bai, S., Kolter, J. Z., Koltun, V., "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling", arXiv preprint arXiv:1803.01271, 2018.

[296]Kingma, D. P., Ba, J., "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, 2014.

Biography

Krešimir Bešenić was born in 1989 in Varaždin, Croatia. He graduated from the Faculty of Electrical Engineering and Computing at the University of Zagreb in 2012 with a master's thesis focused on image processing and machine learning. In 2015, he started working as a research associate at the same institution on a research project funded by Visage Technologies. Currently, he holds the position of a principal research and development engineer in the Face Technology Division at Visage Technologies. His role includes technical leadership on projects related to the development of algorithms and systems for face tracking and analysis, which are also his main scientific interests. He has published papers focused on automatic facial age estimation in relevant scientific journals as well as in the proceedings of international conferences.

List of publications

Journal publications

1. **Bešenić, K.**, Ahlberg, J. and Pandžić, I.S., "Picking out the bad apples: unsupervised biometric data filtering for refined age estimation", The Visual Computer, Vol. 39, Iss. 1, 2023, pp. 219-237, doi:10.1007/s00371-021-02323-y.
2. **Bešenić K.**, Gogić, I., Pandžić, I.S. and Matković, K., "Automatic Image-based Face Analysis Systems Overview", Engineering Power: Bulletin of the Croatian Academy of Engineering, Vol. 13, Iss. 2, 2018, pp. 2-7, uri:https://hrcak.srce.hr/215882.

Conference publications

1. **Bešenić, K.**, Ahlberg, J. and Pandžić, I.S., "Let Me Take a Better Look: Towards Video-Based Age Estimation", in *13th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2024, pp. 57-69, doi:10.5220/0012376800003654.
2. **Bešenić, K**., Ahlberg, J. and Pandžić, I.S., "Unsupervised facial biometric data filtering for age and gender estimation", in *14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 2019, pp. 209-217, doi:10.5220/0007257202090217.

Životopis

Krešimir Bešenić rođen je 1989. godine u Varaždinu. Diplomirao je na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu 2012. godine s magistarskim radom usmjerenim na strojno učenje i obradu slike. 2015. godine počeo je raditi kao znanstveni suradnik u istoj ustanovi na znanstvenom projektu financiranom od strane tvrtke Visage Technologies. Trenutno je na poziciji glavnog inženjera za istraživanje i razvoj u Odjelu za tehnologije lica tvrtke Visage Technologies. Njegova uloga uključuje tehničko vodstvo na projektima vezanim za razvoj algoritama i sustava za praćenje i analizu lica, što su ujedno i njegovi glavni znanstveni interesi. Objavio je radove usredotočene na automatsku procjenu dobi iz lica u relevantnim znanstvenim časopisima kao i u zbornicima međunarodnih konferencija.