

James Z. Wang
Reginald B. Adams, Jr. *Editors*

Modeling Visual Aesthetics, Emotion, and Artistic Style



Modeling Visual Aesthetics, Emotion, and Artistic Style

James Z. Wang • Reginald B. Adams, Jr.
Editors

Modeling Visual Aesthetics, Emotion, and Artistic Style



Editors

James Z. Wang
Information Sciences and Technology
The Pennsylvania State University
University Park, PA, USA

Reginald B. Adams, Jr.
Department of Psychology
The Pennsylvania State University
University Park, PA, USA

ISBN 978-3-031-50268-2

ISBN 978-3-031-50269-9 (eBook)

<https://doi.org/10.1007/978-3-031-50269-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

We dedicate this book to our families, whose love and support have been our constant source of strength and inspiration throughout the entire editing process.

James and Reg

Foreword

Over millennia, people have learned both overt behaviors and more subtle cues to communicate to one another. Anthropological motivations abound, from the need to survive, to protect one's community, and to secure and exploit resources. Along the evolutionary path, people had to interact with others. Even before the emergence of languages, humans had to sense, read, and assign meaning, mostly instantaneously, to observed and perceived facial expressions, body gestures, and actions. Whatever physiological processes were responsible for these abilities, they are now embedded in our genetic makeup and allow children to acquire, learn, and ascribe meaning to social environments.

Over the past century or so, first psychologists, and then computer scientists, began quantitative studies and experiments to try to understand how socially communicative behaviors arose. Fundamental questions about human perception and how sensory systems might be “wired” dominated pre-computational studies. When computers, visual sensors, and displays emerged to support scientific endeavors, new foci arose on linking synthetic perception, algorithmic models of social phenomena, and generalized mechano-robotic and graphical generation of social signals. In other words, the era of synthesized virtual humans began.

I was extremely fortunate to be able to participate in and contribute to this new field from my PhD thesis at the University of Toronto in 1974 up to the present. By 1990, there were several robust research communities interested in human perception, robotics, computer vision, and computer graphics. David Zeltzer of MIT, Brian Barsky of the University of California at Berkeley, and myself from the University of Pennsylvania, organized a “Workshop on the Mechanics, Control and Animation of Articulated Figures” held at the MIT Media Lab in April 1989. We invited participants from multiple perspectives to share their thoughts on humans and their virtual or robotic embodiments. This successful cross-disciplinary meeting led to the first book in 1990—*Making Them Move: Mechanics, Control and Animation of Articulated Figures*—to encompass these disparate but ultimately deeply connected viewpoints.

This volume on *Modeling Visual Aesthetics, Emotion, and Artistic Style*, thoughtfully curated by James Z. Wang and Reginald B. Adams, Jr., is a perfect bookend

to that earlier *Making Them Move* collection. Human behavior observation has been dramatically enabled by low-cost, high-resolution image acquisition hardware feeding real-time computer vision motion analysis systems. Computer virtual human simulation and computer graphics have evolved to the point where non-real-time movie actors, characters, and monsters are produced with efficiency and regularity by the movie and game industry, while real-time human agents are now taking on public roles as announcers, influencers, and assistants. Their ubiquity has fostered artistic interest and study of the human aesthetic. Driving the range of contemporary applications are new tools from the Artificial Intelligence and Machine Learning research communities. We could perhaps only dream about these in 1990. They are the new foundation for human perception and simulation research.

With modern computational tools and decades of computer graphics simulations to build on, additional fascinating aspects of human communication can be studied, modeled, and reproduced. Human emotions, and their companion attributes of mood and personality expressed by face and gesture, have long been of interest to multiple research communities, including the social sciences as well as the computational ones. This volume addresses emotional displays and understanding, including novel dimensions such as threats, which are of clear evolutionary value. As its title aptly describes, this volume also includes new considerations of aesthetics and artistic style. Critical questions of bias and sexual discrimination must be addressed as learning systems depend on datasets that might, inadvertently or naively, perpetuate stereotypes, cultural misconceptions, or prejudices. The maturity of the underlying computational foundations now admits these humanistic questions. A number of works in this volume explore this space of unique human characteristics.

James and Reg have assembled an outstanding collection of current approaches to modeling novel human dimensions. It will be a classic of interdisciplinary computational studies. Enjoy!

Haverford, PA, USA
June 2023

Norman I. Badler

Preface

Visual aesthetics, emotional expression, and artistic style are essential components of human perception and experience, and their significance has only grown with the increasing prevalence of digital media and technologies. The ability to computationally model and analyze these complex concepts has been a longstanding goal in the fields of computer vision, affective computing, and robotics. This timely book represents the collective efforts of active researchers from a diverse set of fields, including computer vision, robotics, psychology, graphics, data mining, machine learning, movement analysis, and art history, who have come together to address these challenging and critical research questions. As our world becomes more interconnected and reliant on digital platforms and artificial intelligence, understanding and effectively utilizing these aspects of human experience has become increasingly important, making this book a vital resource for both researchers and practitioners alike.

The chapters of this book cover a wide range of topics related to the computational modeling of aesthetics, emotion, and artistic style. The first part provides background knowledge related to emotion models and machine learning. The next two parts explore social visual perception in humans and its application to computer vision. Specifically, Part II lays the groundwork by discussing the basic psychological and neurological underpinnings of social and emotional perception from faces and bodies. Part III extends this understanding into the realm of technology, demonstrating methods to train computer systems to detect discrete and micro-momentary emotional expressions from facial and body cues, question the notion of facial neutrality, and broaden the scope of research to include children as well as adults in the context of emotion perception. Part IV focuses on the dynamic intersection of art and technology, shedding light on the language of photography, the interplay between breath-driven robotic performances and human dance, and the application of machine learning in the contextual analysis of artistic style. The remaining three parts dive deeper into the computational modeling of visual aesthetics, emotion, and artistic style.

One of the unique features of this book is its multidisciplinary approach, bringing together contributions from various domains, such as computer science, psychology,

art history, and cognitive science. This interdisciplinary approach fosters a more holistic understanding of the subject matter and encourages cross-disciplinary collaboration, leading to novel insights and advancements in the field.

The versatile nature of the book format has enabled us to encompass an array of contribution types. These include comprehensive tutorials and reviews of theoretical frameworks and computational methodologies, extensive literature surveys, novel methodological approaches, in-depth case studies, insightful opinion pieces, rigorous empirical investigations, and comparative analyses.

Another feature of this book is its focus on cutting-edge research. The methods and information presented in this book represent the latest developments in the field and have the potential to significantly advance the field. The comprehensive and in-depth treatment of topics offered by book chapters provides a richer understanding of the subject for readers, which can be especially beneficial for those new to a new interdisciplinary field or those looking to expand their knowledge.

Finally, the impact of the results presented in the book can be far-reaching. The ability to computationally model aesthetics, emotion, and artistic style has the potential to enable many computer and robotic applications that can benefit millions of people around the world. From children needing care to the elderly needing assistance, from amateur photographers to people working alongside robots, the impact of this work is broad.

We trust that this book will serve as a valuable resource for researchers, practitioners, educators, and students who are interested in advancing the field. The cross-disciplinary nature of the book increases the chances of a wider audience accessing the research, leading to broader dissemination and long-term recognition of the presented findings. We hope that this book will inspire further research, foster interdisciplinary collaboration, and contribute to the advancement of computational modeling of visual aesthetics, emotion, and artistic style.

State College, PA, USA
June 2023

James Z. Wang
Reginald B. Adams, Jr.

Acknowledgments

First and foremost, we would like to express our deepest gratitude to our colleagues who contributed their invaluable expertise, knowledge, and insights to this book. Their dedication and commitment to advancing this field have substantially enriched the content and elevated the overall quality of this work. It has been a privilege to collaborate with such an accomplished and diverse assembly of researchers in this endeavor.

We convey our sincere appreciation to the editorial team at Springer Nature for their steadfast support, professionalism, and guidance throughout the publication process. Their invaluable feedback and constructive suggestions have played an important role in shaping the final product. Special thanks go to our editor, Susan E. Grove, and project coordinator, Arun S. Shanmugam, for their relentless enthusiasm and encouragement, which motivated us to strive for excellence in our work. Additionally, we would like to express our gratitude to the anonymous reviewers for their valuable insights and constructive feedback on our book proposal.

We deeply appreciate Norman Badler for his insightful Foreword. His rich experiences have helped contextualize the evolution of our field, emphasizing the roles of artificial intelligence and machine learning in human perception and simulation research. His emphasis on addressing issues like bias and cultural misconceptions in learning systems is invaluable.

We would also like to thank our academic mentors, advisees, colleagues, and collaborators who have inspired and supported our research over the years. Their expertise, encouragement, and friendship have been essential in the development of our understanding and passion for the field. In particular, J. Z. Wang is grateful to Gio Wiederhold, Dennis A. Hejhal, Martin A. Fischler, and Edward H. Shortliffe for their invaluable guidance, wisdom, and belief in his potential. He is also grateful for the support and encouragement received from Adam Fineberg, Yelin Kim, Tatiana D. Korelsky, and Juan P. Wachs. R. B. Adams, Jr. is particularly grateful to Robert E. Kleck, Ursula Hess, and Nalini Ambady for their early mentorship and encouragement. He is also grateful to all his colleagues in psychology and vision science who have helped him champion the field of Social Vision. Finally, he is grateful to J. Z. Wang for spearheading this book, for collaboration over the years

extending his own work in new directions, and enabling him to apply insights from social visual perception to this burgeoning field of computer vision.

Our deepest appreciation goes to the countless artists and creators whose work has inspired and fueled our research. Their artistic expressions and creative pursuits provide the foundation for our exploration of this fascinating field.

We gratefully acknowledge the financial support provided by the National Science Foundation (NSF) under Grant Nos. 1110970, 1921783, and 2234195, which has been instrumental in advancing the research presented in this book. This funding has enabled us to pursue innovative research directions, collaborate with leading experts, and ultimately contribute to the growing body of knowledge in this field. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. Additionally, J. Z. Wang extends his sincere appreciation to the Amazon Research Awards program for their gift, which has further bolstered the research of his team and facilitated the development of cutting-edge methodologies and technologies. His team's research in visual art has been supported in part by the National Endowment for the Humanities (NEH) under Grant Nos. HAA-271801-20 and HAA-287938-22. His team's research in machine learning has been supported in part by the NSF under Grant Nos. 2205004 and 2216127. We are immensely grateful for the confidence and investment these organizations have placed in our work, and we strive to continue making meaningful contributions to this exciting and dynamic field.

Last but not least, we would like to express our heartfelt gratitude to our family and friends for their constant love, support, and encouragement. Their belief in our abilities, patience with our countless hours of work, and persistent understanding have been the bedrock of our endeavors. Without them, this book would not have been possible. J. Z. Wang would like to especially thank Jia Li and their children Justina and Nora Wang for their inspiration and consistent understanding. Similarly, R. B. Adams, Jr. thanks Katharine Donnelly Adams as well as their children Henry and Lena Logan Adams for their understanding and encouragement throughout the process of putting this book together.

Contents

Part I Foundations of Emotion Modeling and Machine Learning

- | | | |
|----------|--|-----------|
| 1 | Models of Human Emotion and Artificial Emotional Intelligence | 3 |
| | Benjamin Wortman | |
| 2 | A Concise Introduction to Machine Learning | 23 |
| | Sitao Zhang | |

Part II Human Social Vision

- | | | |
|----------|---|-----------|
| 3 | Facing a Perceptual Crossroads: Mixed Messages and Shared Meanings in Social Visual Perception | 45 |
| | Natalie Strand, Nicole Hedgecoth, and Reginald B. Adams, Jr. | |
| 4 | Social Vision of the Body in Motion: Interactions Between the Perceiver and the Perceived | 59 |
| | Pamala N. Dayley and Kerri L. Johnson | |
| 5 | Visual Perception of Threat: Structure, Dynamics, and Individual Differences | 71 |
| | Kestutis Kveraga | |
| 6 | From Pixels to Power: Critical Feminist Questions for the Ethics of Computer Vision | 91 |
| | Flora Oswald | |

Part III Computer Social Vision

- | | | |
|----------|---|------------|
| 7 | High-Speed Joint Learning of Action Units and Facial Expressions .. | 105 |
| | Feng Xu, Yifan Yuan, Junping Zhang, and James Z. Wang | |
| 8 | ExpressionFlow: A Microexpression Descriptor for Efficient Recognition | 127 |
| | Feng Xu, Yifan Yuan, Junping Zhang, and James Z. Wang | |

9 Emotion in the Neutral Face: Applications for Computer Vision and Aesthetics	147
Daniel N. Albohn and Joseph C. Brandenburg	
10 Multi-Stream Temporal Networks for Emotion Recognition in Children and in the Wild	163
Panagiotis P. Filntisis, Niki Efthymiou, Gerasimos Potamianos, and Petros Maragos	

Part IV Photography, Arts

11 The Formal Language of Photography: A Primer	181
QT Luong	
12 Breathing with Robots: Notating Performer Strategy, Alongside Choreographer Intent and Audience Observation, in Breath-Driven Robotic Dance Performance	203
Kate Ladenheim, Amy LaViers, and Catherine Maguire	
13 Humanist-in-the-Loop: Machine Learning and the Analysis of Style in the Visual Arts	219
Kathryn Brown	

Part V Aesthetics

14 The Inter-Relationship Between Photographic Aesthetics and Technical Quality	231
Franz Götz-Hahn, Lai-Kuan Wong, and Vlad Hosu	
15 Image Restoration for Beautification	257
Dejia Xu, Yifan Jiang, and Zhangyang Wang	
16 Image Affect Modeling: An Industrial Perspective	279
Xin Lu	

Part VI Emotion

17 Emotional Expression as a Means of Communicating Virtual Human Personalities	293
Sinan Sonlu, Khasmamat Shabanovi, Uğur Güdükbay, and Funda Durupinar	
18 Modeling Emotion Perception from Body Movements for Human-Machine Interactions Using Laban Movement Analysis	313
Tal Shafir	
19 Demographic Differences and Biases in Affect Evoked by Visual Features	331
Baris Kandemir, Hanjoo Kim, Michelle G. Newman, Reginald B. Adams, Jr., Jia Li, and James Z. Wang	

Part VII Artistic Style

20 Deep Network-Based Computational Transfer of Artistic Style in Art Analysis.....	351
David G. Stork	
21 Balance of Unity and Variety in Fine Art Paintings: A Computational Study	369
Jia Li	
Index.....	393

Contributors

Reginald B. Adams, Jr. is a Professor of Psychology at The Pennsylvania State University. He received his Ph.D. in Social Psychology from Dartmouth College in 2002. Reg is interested in how we extract social and emotional meaning from nonverbal cues, particularly via the face. His work addresses how multiple social messages (e.g., emotion, gender, race, age, etc.) combine across multiple modalities and interact to form the unified representations that guide our impressions of and responses to others. Although his questions are social psychological in origin, his research draws upon vision cognition and affective neuroscience to address social perception at the functional and neuroanatomical levels. With his colleagues, Reg helped establish and champion the subfield of Social Vision by publishing an edited volume titled *The Science of Social Vision* (Adams, Ambady, Nakayama, & Shimojo, 2010, Oxford University Press). His research has been funded by NSF, NIA, and NIMH (NIH).

Daniel N. Albohn is a Principal Researcher at the University of Chicago Booth School of Business. Dan received his Ph.D. in Psychology from The Pennsylvania State University. His research uses data-driven, machine learning, and human responses to examine how social cues inform judgments of people and objects. He has a particular interest in how individuals extract information from neutral or minimally expressive faces.

Norman I. Badler is an Emeritus Professor of Computer and Information Science at the University of Pennsylvania. He received his B.A. in Creative Studies Mathematics from the University of California Santa Barbara in 1970, his M.S. in Mathematics from the University of Toronto in 1971, and his Ph.D. in Computer Science from the University of Toronto in 1975. His research has involved developing software to acquire, simulate, animate, and control 3D computer graphics human body, face, gesture, locomotion, and manual task motions, both individually and for heterogeneous groups. These virtual humans are meant to portray physical, cognitive, perceptual, personality, relationship, and cultural parameters. He has

supervised or co-supervised 61 Ph.D. students, many of whom have become academics or researchers in the movie visual effects and game industries. He was the founding Director of the SIG Center for Computer Graphics, the Center for Human Modeling and Simulation, and the ViDi Center for Digital Visualization at Penn. He has co-authored five books, one in digital human modeling and the other four in virtual crowd simulation. He serves part-time as the Head of Metaverse Research at Philadelphia-based Cesium GS Inc.

Joseph C. Brandenburg is a fourth-year graduate student in the school psychology program at The Pennsylvania State University. He has been working in and collaborating with the members of the Social Vision and Interpersonal Perception lab for 9+ years throughout his undergraduate and postgraduate career. Joe has his Master of Science in Clinical Psychology from Millersville University and his Master of Education in School Psychology. Joe's research interests include emotion perception, emotion regulation, stress, and psychophysiology. He has worked on myriad projects including these topics with an emerging interest in how wearable technologies can help within these already existing arenas of interest.

Kathryn Brown is an Associate Professor of Art History at Loughborough University (UK). Her books include *Women Readers in French Painting 1870–1890* (2012), *Matisse's Poets: Critical Performance in the Artist's Book* (2017), *Henri Matisse* (2021), and *Dialogues with Degas: Influence and Antagonism in Contemporary Art* (2023). She has edited several essay collections, including *Digital Humanities and Art History* (Routledge, 2020). Brown's research has been supported by numerous funders including the Association of Art History (UK), the British Academy, the Independent Social Research Foundation, and the Terra Foundation for American Art. In 2021, Brown was a Paul Mellon Visiting Senior Fellow at the Center for Advanced Study in the Visual Arts (Washington, DC).

Pamala N. Dayley is a third-year Social Psychology graduate student at UCLA, supervised by Dr. Kerri Johnson. She received her Bachelor's degree from The Pennsylvania State University, Abington campus, and her Master's degree from UCLA. Her research interests include the perceptions of others (face and body), judgments made about targets based on perceptual cues, and the downstream consequences (e.g., discrimination) of said judgments. She is a National Science Foundation awardee and a National Defense Science and Engineering Graduate Fellow.

Funda Durupinar received a B.S. degree from Middle East Technical University, Ankara, Turkey and an M.S. degree from Bilkent University in Computer Engineering in 2002 and 2004, respectively. She received her Ph.D. degree in August 2010 from the Department of Computer Engineering at Bilkent University, Ankara, Turkey. After completing her Ph.D., she worked as a Postdoctoral Researcher at the Center for Human Modeling and Simulation, University of Pennsylvania. She

worked as a Software Engineer at Memorial Sloan Kettering Cancer Center and as a Senior Research Associate at Oregon Health & Science University. She is currently an Assistant Professor in the Department of Computer Science at the University of Massachusetts at Boston. Her research links computer graphics, artificial intelligence, and psychology with a focus on creating believable virtual humans.

Niki Efthymiou is a Ph.D. student at the School of Electrical and Computer Engineering, National Technical University of Athens (NTUA), under the supervision of Prof. Petros Maragos. She is working primarily in computer vision problems associated during Human–Robot Interaction. She is a Researcher at the Computer Vision, Speech Communication, and Signal Processing Group at NTUA, and her research interests lie in the fields of gesture, action, and emotion recognition, with a focus on Child–Robot Interaction. She received her Diploma degree in Applied Mathematics and Master’s degree in Computational Mechanics from NTUA.

Panagiotis P. Filntisis is a Postdoctoral Researcher at the IRAL lab of the National Technical University of Athens and a Research Assistant at the Athena Research and Innovation Center. He received his Ph.D. in 2022 under the supervision of Prof. Petros Maragos and holds an M.Eng. Diploma degree in ECE from NTUA. His work lies at the crossroads of computer vision and audio processing for affective computing.

Franz Götz-Hahn is currently working as a Postdoctoral Researcher in the Intelligent Embedded Systems group at the University of Kassel, where he heads the AI for Motion research group. He received his M.Sc. in Artificial Intelligence from Maastricht University, Netherlands and the Ph.D. in Computer Science from the University of Konstanz, Germany with his thesis titled “Video Quality Assessment in-the-wild.” Franz’s dissertation was the culmination of pioneering work in the field of deep learning for image and video quality assessment, including the (co-)authorship of KonVid-1k and KonVid-150k, two of the most influential and largest in-the-wild video quality datasets to date. Recently, he has expanded his expertise beyond image and video quality toward using artificial intelligence more generally in domains involving motion, such as in automotive.

Uğur Güdükbay received a B.S. degree in Computer Engineering from the Middle East Technical University, Ankara, Turkey, in 1987 and an M.S. and Ph.D. degrees in Computer Engineering and Information Science from Bilkent University, Ankara, Turkey, in 1989 and 1994, respectively. He conducted research as a Postdoctoral Fellow at the Human Modeling and Simulation Laboratory at the University of Pennsylvania. Currently, he is a Professor in the Department of Computer Engineering at Bilkent University. His research interests include human modeling and animation, conversational virtual agents, personality and emotion synthesis,

crowd simulation, rendering, and visualization. He is a senior member of ACM and IEEE.

Nicole Hedgecoth is a doctoral candidate at The Pennsylvania State University. They earned their Bachelor's degree in Psychology and their first Master's degree in Negotiation and Conflict management from the University of Baltimore, before continuing to earn their second Master's in Psychology at The Pennsylvania State University. Nikki's research interest is in bringing an interdisciplinary approach to social vision and person perception, drawing on feminist theory and methods to inform their work.

Vlad Hosu is a Postdoctoral Researcher in the Multimedia Signal Processing group at the University of Konstanz. With a Ph.D. in Computer Vision from the National University of Singapore, his dissertation focused on the aesthetics of lighting design in computational photography. Vlad is dedicated to exploring the intersection of technical and aesthetic quality assessment by studying human visual perception. He is developing innovative visual quality models using machine-learning techniques and crowdsourcing. Vlad's contributions include co-authoring several central databases for the field and widely used predictive models.

Yifan Jiang is a Ph.D. student in the Department of Electrical and Computer Engineering at the University of Texas at Austin, supervised by Prof. Zhangyang (Atlas) Wang. He received his Bachelor's degree from Huazhong University of Science and Technology, Wuhan, China. His research interests range from neural rendering, generative models, and computational photography. He also completed internships at Bytedance AI Lab, Adobe, and Google Research. He is a recipient of the 2023 Apple Scholar in AI/ML.

Kerri L. Johnson is a Professor in the departments of Communication and Psychology at UCLA, where she currently serves as the Associate Vice Chancellor for Faculty Development. After receiving her Ph.D. from Cornell University in 2004, she was a Postdoctoral Fellow at NYU before joining the faculty at UCLA. Her research is at the forefront of the burgeoning field of Social Vision. She examines how the perception of cues in the face and body impacts interpersonal judgments, behaviors, and biases. She has published widely on how such perceptions inform a range of downstream judgments, including politics, sexual orientation, and even religion, often documenting profound biases in how appearance impacts meaningful outcomes. Her lab pursues highly interdisciplinary work, including both theoretical and methodological breadth from across the allied social, cognitive, and visual sciences.

Baris Kandemir received a B.Sc. degree in Electrical and Electronics Engineering from Boğaziçi (Boğaziçi) University, Istanbul, Turkey with high honor in 2012. He obtained his Ph.D. degree from the College of Information Sciences and

Technology, The Pennsylvania State University, University Park, in 2019. Since July 2018, he has been with DeepMap, a subsidiary of NVIDIA, where he is a systems software engineer. His main interests are biomedical image processing, computational aesthetics and affect, and 3D computational geometry. During his Ph.D. studies, he investigated visual balance and aesthetics, relationship among visual cues, affect, and demographics. Additionally, he studied the 3D segmentation of tubular structures in confocal microscopy stacks. He is currently working on crowdsourced mapping using NVIDIA's perception modules.

Hanjoo Kim is a Postdoctoral Research Fellow at the Heinz C. Prechter Bipolar Research Program, University of Michigan. He obtained a Ph.D. in Clinical Psychology from The Pennsylvania State University and completed an APA-accredited psychology internship at the New Mexico VA/Southwest Consortium. His primary research interests center around the “underlying mechanisms” of emotional disorders, including anxiety, unipolar depression, and bipolar spectrum disorders. Currently, his research focuses on understanding the emotion dysregulation processes involved in repetitive negative thoughts, such as worry and rumination. To investigate this topic, he is utilizing various psychophysiological methodologies, such as skin conductance, emotional facial expressions, and heart rate variability, alongside intensive longitudinal data analysis.

Kestutis Kveraga is an Assistant Professor at the Harvard Medical School and an Assistant in Neuroscience at the Massachusetts General Hospital. He is a cognitive neuroscientist who studies the neural mechanisms of threat perception from naturalistic stimuli, with strong interests in visual pathway function and autism. He is also interested in neural aesthetics and how brain activity can be employed to predict and shape architectural design and art. He has expertise in neuroimaging methods, such as structural and functional MRI (including ultra-high-field high-resolution 7T fMRI), MEG and EEG, psychophysical techniques (eye and limb tracking, visual pathway biasing), and brain connectivity analyses (e.g., Dynamic Causal Modeling and biomagnetic phase synchrony).

Kate Ladenheim is a choreographer, media designer, and creative technologist who researches bodies in motion and how they impact and are impacted by systems of social and technological pressure. Her work has been presented internationally and spans interactive installations, media design, dance performance, and robotics research. Ladenheim holds an M.F.A. in Media Design Practices from ArtCenter College of Design. She recently assisted robotics research at UCLA, and was the 2019–2020 Artist in Residence at the Robotics, Automation, & Dance (RAD) Lab at UIUC. Her work was celebrated in *Dance Magazine* as one of their “25 to Watch” and “Best of 2018.” She is the current Artist in Residence at the Maya Brin Institute for New Performance, a faculty role at the University of Maryland—College Park.

Amy LaViers is the Director of the Robotics, Automation, and Dance (RAD) Lab. Her choreography and machine designs have been presented internationally, includ-

ing at Joe’s Pub at the Public Theater and the Performance Arcade. Her writing has appeared in academic journals like *Nature* and *Robotics and Autonomous Systems* as well as public venues like American Scientist and Aeon. She is a recipient of DARPA’s Young Faculty Award (YFA), and her teaching has been recognized on the list of Teachers Ranked as Excellent by Their Students, with outstanding distinction, at the University of Illinois at Urbana-Champaign (UIUC). She has held positions as a co-founder of three start-up companies and as an engineering faculty member at UIUC and the University of Virginia. She holds a CMA from the Laban/Bartenieff Institute of Movement Studies, a Ph.D. and M.S. from Georgia Institute of Technology, and a B.S.E. and certificate in Dance from Princeton University.

Jia Li is a Professor of Statistics and (by courtesy) Computer Science at The Pennsylvania State University. Her research interests include machine learning and image analysis. For her innovations in image retrieval, annotation, aesthetics/composition analysis, and other areas, she has been awarded sixteen US patents. She worked as a Program Director at the National Science Foundation from 2011 to 2013, a Visiting Scientist at Google Labs in Pittsburgh from 2007 to 2008, and a Researcher at the Xerox Palo Alto Research Center from 1999 to 2000. She received an M.Sc. degree in Electrical Engineering (1995), an M.Sc. degree in Statistics (1998), and a Ph.D. degree in Electrical Engineering (1999) from Stanford University. She was Editor-in-Chief of *Statistical Analysis and Data Mining: The ASA Data Science Journal* from 2018 to 2020. She is a Fellow of the Institute of Electrical and Electronics Engineers and a Fellow of the American Statistical Association.

Xin Lu received her Ph.D. degree from the College of Information Sciences and Technology, The Pennsylvania State University, University Park in 2016. Prior to that, she received a B.E. and B.A. degrees in Electronic Engineering and English and an M.E. degree in Signal and Information Processing, all from Tianjin University, China. Since August 2015, she has been with Adobe Inc., where she is currently a Senior Manager and Scientist. Her main research interests are image generation, image segmentation, image aesthetics and emotions, and efficient neural networks. During her Ph.D. studies, she discovered and verified the relationship between simplicity and valence and angularity and valence in complex scenes.

QT Luong is a former computer vision researcher with positions at the University of California, Berkeley, and SRI International turned freelance photographer. His Ph.D. thesis, “Fundamental Matrix and Self-calibration,” introduced concepts that spawned a decade of research. The resulting 1992 European Conference on Computer Vision paper “Camera self-calibration: Theory and experiments” (with O. Faugeras and S. Maybank) won the inaugural Koenderink Prize for Fundamental Contributions in Computer Vision in 2008. He is the coauthor (with O. Faugeras) of the book *The Geometry of Multiple Images* (MIT Press 2001/2004). Luong was the first to photograph all of America’s 63 national parks—in large format.

He received the Sierra Club's Ansel Adams Award for Photography and the National Parks Conservation Association's Robin W. Winks Award for Enhancing Public Understanding of National Parks. His best-selling book *Treasured Lands: A Photographic Odyssey Through America's National Parks* (2016) won 12 national and international book awards.

Catherine Maguire is a movement educator and dance artist. She is a master teacher of the Laban/Bartenieff Movement System (LBMS) and a Certified Movement Analyst (CMA), having taught and co-coordinated movement analysis certification training programs in the USA, Europe, Mexico, and China. Maguire is a faculty member of WholeMovement, a coterie of movement analysts working together to promote movement studies globally. She has coauthored several publications on expressive robotic systems, including *Making Meaning with Machines: Somatic Strategies, Choreographic Technologies and Notational Abstractions Through a Laban/Bartenieff Lens*. She was the Founder and Artistic Director of Offspring Dance Company in New York City and the Founder and Head of the dance program at Drew University in Madison, NJ, as well as Assistant Professor of dance at Piedmont Virginia Community College. She lives in central Virginia where she teaches ongoing movement classes designed to foster self-expression, body connectivity, and transformation through movement.

Petros Maragos is a full Professor of the School of Electrical and Computer Engineering, National Technical University of Athens, Greece, and Director of the Intelligent Robotics and Automation Lab and the CVSP Group. He has worked as a Professor at USA universities, including Harvard University (1985–93) and Georgia Tech (1993–98). He is also the coordinator of a Robotics Research Unit at the Athena Research and Innovation Center. His research and teaching interests include signal processing and machine learning, computer vision, speech/language, and robotics. He is the recipient of several awards for his academic work. He has been the PI of several US, European, and Greek research projects and served as General Chair of EUSIPCO'17 and ICASSP'23. He is a Fellow of IEEE and EURASIP.

Michelle G. Newman is a Professor of Psychology and Psychiatry and Director of the Center for the Treatment of Anxiety and Depression at The Pennsylvania State University. She received her Ph.D. in Clinical Psychology from the University of Stony Brook in 1992 and completed a postdoctoral fellowship at Stanford University in 1994. Dr. Newman has conducted basic and applied research on anxiety disorders and depression and has published over 200 papers on these topics. She is the past editor of *Behavior Therapy* and is currently Associate Editor of the *Journal of Anxiety Disorders*. She is also the recipient of the APA Division 12 Turner Award for distinguished contribution to clinical research, APA Division 29 Award for Distinguished Publication of Psychotherapy Research Award, ABCT Outstanding Service Award, APA Division 12 Toy Caldwell-Colbert Award for Distinguished

Educator in Clinical Psychology, and Raymond Lombra Award for Distinction in the Social or Life Sciences. She is also a Fellow of the American Psychological Association Divisions 29 and 12, the Association for Behavioral and Cognitive Therapies, and the American Psychological Society.

Flora Oswald is a recent graduate from Penn State's dual-title doctoral program in Psychology and Women's, Gender, and Sexuality Studies, and an Assistant Research Professor at the University of Connecticut. Flora is interested in how marginalized identities shape people's experiences and perceptions of their social worlds, with a particular focus on stereotyping and stigmatization. Much of Flora's current work bridges feminist social psychological approaches with visual perception research to elucidate a feminist social vision perspective that prioritizes marginalized perceivers. Flora's work has been supported by awards from the Social Sciences and Humanities Research Council of Canada, Women and Gender Equality Canada, the Government of Alberta, and the Society for Personality and Social Psychology, among others.

Gerasimos Potamianos is an Associate Professor in the Department of Electrical and Computer Engineering at the University of Thessaly in Greece and holds a Ph.D. degree from Johns Hopkins University (1994). Prior to his current position, he has been at the Center for Language and Speech Processing at Johns Hopkins, at AT&T Labs-Research, and the IBM T.J. Watson Research Center in the USA, followed by FORTH and Demokritos Research Centers in Greece. His research interests span multisensory and multimodal speech processing and scene analysis with applications to human-computer/robot interaction and ambient intelligence. He has authored 160 articles that have received over 6.5k citations, holds 7 patents, and has been involved in numerous European and national research projects. He has served as an organizing committee member of EUSIPCO'17, SLT'18, and ICASSP'23, at the IEEE Speech and Language Committee, and is currently a member of IEEE, ISCA, EURASIP, and the Technical Chamber of Greece.

Khasmamat Shabanovi received his B.S. degree in Computer Engineering from Bilkent University. During his senior year, he researched discovering a correlation between pose and apparent personality traits. Currently, he is pursuing an M.S. degree in Computer Science at the Technical University of Munich. His research interests are artificial intelligence and deep learning.

Tal Shafir graduated from law school at the Hebrew University of Jerusalem. Following her passion, she then studied dance-movement therapy at the University of Haifa, and completed her Ph.D. in neurophysiology of motor control and two postdoctoral fellowships: in brain-behavior interactions in infants, and in affective neuroscience, all at the University of Michigan. While working as a research investigator at the University of Michigan, Department of Psychiatry, she started to develop her research on movement-emotion interaction and its underlying brain

mechanisms, behavioral expressions, and therapeutic applications, which she now continues at the University of Haifa. Shafir, certified also in Laban Movement Analysis, was the main editor of *The Academic Journal of Creative Arts Therapies*, and of *Frontiers in Psychology* research topic: “The state of the art in creative arts therapies.” She has been serving on The American Dance Therapy Association (ADTA) research committee since 2016, and was the recipient of ADTA 2020 Innovation Award.

Sinan Sonlu received his B.S. and M.S. degrees in Computer Engineering from Bilkent University. He is currently pursuing Ph.D. studies at the same university. His research interests include conversational virtual agents, expressive animation, motion generation, and personality synthesis. With his research group, he currently works on successfully representing the different personality traits in virtual character animation. He and his colleagues recently published their conversational agent framework with multimodal personality expression.

David G. Stork is an Adjunct Professor at Stanford University. He received a B.S. degree in physics from the Massachusetts Institute of Technology, Cambridge, MA, USA in 1976 and a Ph.D. degree in physics from the University of Maryland, College Park, MD, USA in 1984. He has made contributions to machine learning, pattern recognition, computer vision, artificial intelligence, computational optics, image analysis of fine art, and related fields. He is a Fellow of seven international scholarly societies, and his eight books/proceedings volumes include the second edition of *Pattern Classification* and *Pixels and Paintings: Foundations of Computer-Assisted Connoisseurship*.

Natalie Strand is a graduate student at The Pennsylvania State University working with Dr. Reg Adams. She has a B.S. in Behavioral and Cognitive Neuroscience and is currently pursuing a Ph.D. in Psychology with a specialization in cognitive and affective neuroscience. Natalie’s research interests are broadly related to the influence of compound social cues on emotion perception in the face. More specifically, her work bridges social vision and feminist approaches to investigate how structural facial cues and gender/sex emotion stereotypes influence emotion perception within and outside of the gender binary. Natalie is also interested in exploring the use of computer vision models to examine how facial cues impact emotion perception.

Zhangyang Wang is currently the Jack Kilby/Texas Instruments Endowed Assistant Professor of Electrical and Computer Engineering at The University of Texas at Austin. He received his Ph.D. degree in ECE from UIUC in 2016, advised by Professor Thomas S. Huang; and his B.E. degree in EEIS from USTC in 2012. Prof. Wang has broad research interests spanning from the theory to the application aspects of machine learning (ML). At present, his core research mission is to leverage, understand, and expand the role of sparsity, from classical optimization

to modern neural networks, whose impacts span many important topics such as efficient training/inference/transfer, robustness and trustworthiness, generative AI, and graph learning. Prof. Wang has received many research awards, including an NSF CAREER Award, an ARO Young Investigator Award, an IEEE AI's 10 To Watch Award, an INNS Aharon Katzir Young Investigator Award, and a few more industry research awards.

James Z. Wang is a Distinguished Professor at The Pennsylvania State University. He received a Bachelor's degree in Mathematics *summa cum laude* from the University of Minnesota (1994) and an M.S. degree in Mathematics (1997), an M.S. degree in Computer Science (1997), and a Ph.D. in Medical Information Sciences (2000), all from Stanford University. His research interests include affective computing, image analysis, image modeling, image retrieval, and their applications. He was a Visiting Professor at the Robotics Institute at Carnegie Mellon University (2007–2008), a lead special section Guest Editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2008), and a program manager at the Office of the Director of the National Science Foundation (2011–2012). He is also affiliated with the Department of Communication and Media, School of Social Sciences and Humanities, Loughborough University, UK (2023–2024).

Lai-Kuan Wong is currently an Associate Professor with the Faculty of Computing and Informatics and the Chair of the Center for Visual Computing at the Multimedia University, Malaysia. She received the B.Sc. degree in Computer Science from Universiti Sains Malaysia, Malaysia and the M.Sc. and Ph.D. degrees in Computer Science from the National University of Singapore. Her research interests include computational photography, computational aesthetics, stereo image and video enhancement, and medical imaging. She serves as Co-Chair for several international workshops held in conjunction with Asian Conference on Pattern Recognition 2015, Asian Conference on Computer Vision 2018, and ACM Multimedia 2020, and as the Organizing Committee for several international conferences including International Conference on Image Processing 2023, ACM Multimedia Asia 2023, IEEE International Conference on Multimedia & Expo 2022, International Symposium on Intelligent Signal Processing and Communication Systems 2022, and Workshop on Multimedia Signal Processing 2018.

Benjamin Wortman is a Ph.D. candidate in the Informatics program at The Pennsylvania State University. He received a Bachelor's degree in Data Sciences and an M.S. degree in Informatics from The Pennsylvania State University, University Park, in 2020 and 2022, respectively. His research interests include affective computing, computer vision, and machine learning.

Dejia Xu is a Ph.D. student from the Department of Electrical and Computer Engineering at the University of Texas at Austin, advised by Prof. Zhangyang (Atlas) Wang. He received his B.S. degree from the School of Electronics Engineering

and Computer Science at Peking University in 2021. His research interests include computational photography, creative vision, and implicit neural representation. He is one of the recipients of the 2022 Snap Research Fellowship.

Feng Xu received a B.S. degree and an M.S. degree in Computer Science from Fudan University in 2013 and 2016, respectively. His research interests include machine learning, computer vision, and affective computing. He has been engaged in cyber risk management at Ant Financial Group since 2016.

Yifan Yuan received a B.S. degree in Physics from Fudan University in 2019 and is now a Ph.D. student in Computer Science at Fudan University admitted in 2021. Her research interests include image attribute manipulation, microexpression recognition, and image generation.

Junping Zhang received a B.S. degree in Automation from Xiangtan University, China, in 1992, an M.S. degree in Control Theory and Control Engineering from Hunan University, Changsha, China, in 2000, and a Ph.D. degree in Intelligent Systems and Pattern Recognition from the Institution of Automation, Chinese Academy of Sciences, in 2003. He has been a Professor at the School of Computer Science, Fudan University, since 2006. His research interests include machine learning, image processing, biometric authentication, and intelligent transportation systems. He has been an Associate Editor of *IEEE Intelligent Systems* since 2009 and was an Associate Editor of *IEEE Transactions on Intelligent Transportation Systems* (2010–2018).

Sitao Zhang is a Ph.D. candidate in the Informatics program at The Pennsylvania State University, advised by James Z. Wang. His primary focus of research centers on the fields of computer vision and machine learning, with a specific emphasis on vision-language integration and self-supervised learning. Before joining Penn State, he received a Master’s degree in Mathematics from the University of Wisconsin-Madison and a Bachelor’s degree in Statistics from Sun Yat-sen University.

Acronyms

AAM	Active Appearance Model
ACG	Attributed Composition Graph
ACLU	American Civil Liberties Union
ACM	Association for Computing Machinery
ADTA	American Dance Therapy Association
AEI	Artificial Emotional Intelligence
AI	Artificial Intelligence
AMT	Amazon Mechanical Turk
ANE	Apple Neural Engine
ANOVA	Analysis of Variance
APA	American Psychological Association
ARO	Army Research Office
ASA	American Statistical Association
ASD	Autism Spectrum Disorders
ASM	Active Shape Model
AU	Action Unit
AUC	Area Under the Curve
AUC ROC	Area Under the Receiver Operating Characteristic Curve
BEEU	Bodily Expressed Emotion Understanding
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
BoLD	Body Language Dataset
CMA	Certified Movement Analyst
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CRI	Child-Robot Interaction
CT-MC	Continuous Time Markov Chain
DARPA	Defense Advanced Research Projects Agency
DCT	Discrete Cosine Transform
DNN	Deep Neural Network
DSLR	Digital Single-Lens Reflex
DT-MC	Discrete Time Markov Chain

EDR	Endpoint Detection and Response
EEG	Electroencephalography
EMFACS	Emotion Facial Action Coding System
FACS	Facial Action Coding System
fMRI	Functional Magnetic Resonance Imaging
FPS	Frame-per-Second
GAN	Generative Adversarial Network
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HDR	High Dynamic Range
HEIC	High Efficiency Image Container
HICEM	High-Coverage Emotion Model
HRI	Human-Robot Interaction
HSI	Hue, Saturation, Intensity
I/O	Input and Output
IAA	Image Aesthetics Assessment
IQA	Image Quality Assessment
IEEE	Institute of Electrical and Electronics Engineers
ISP	Image Signal Processor
JPEG/JPG	Joint Photographic Experts Group
LBMS	Laban/Bartenieff Movement System
LDA	Latent Dirichlet Allocation
LMA	Laban Movement Analysis
MEG	Magnetoencephalography
ML	Machine Learning
MP	McCulloch-Pitts
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NIQE	Natural Image Quality Evaluator
NLP	Natural Language Processing
NSF	National Science Foundation
OCEAN	Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism
OEM	Original Equipment Manufacturer
PAD	Pleasure, Arousal, and Dominance
PCA	Principal Component Analysis
PLD	Point-Light Display
PSNR	Peak Signal-to-Noise Ratio
RAM	Random-Access Memory
RBF	Radial Basis Function
RGB	Red, Green, and Blue
RYB	Red, Yellow, and Blue
SES	Socio-Economic Status
SVM	Support Vector Machine
SVR	Support Vector Regression

TIPI	Ten Item Personality Inventory
TSN	Temporal Segment Network
UI/UX	User Interface and User Experience
VAD	Valence, Arousal, and Dominance
VR	Virtual Reality

Part I

Foundations of Emotion Modeling and Machine Learning

Because this book is multidisciplinary in nature, this part will provide essential knowledge on emotion models and machine learning fundamentals.

Chapter 1, “Models of Human Emotion and Artificial Emotional Intelligence,” aims to bridge the gap between emotion models used in psychology and their application in affective computing tasks. It surveys existing emotion models in psychology, highlighting their strengths and weaknesses for computational tasks involving human emotion.

Chapter 2, “A Concise Introduction to Machine Learning,” offers a fundamental understanding of machine learning techniques relevant to the theme of the book. This chapter serves as a starting point for readers with limited or no relevant expertise. It introduces learning algorithms, basic concepts, and fundamental principles in machine learning systems.

These chapters serve as a valuable foundation for readers interested in emotion modeling and the application of machine learning in aesthetics, emotion, and artistic style. By familiarizing themselves with emotion models and understanding the basics of machine learning, readers can better comprehend and engage with recent research in these areas presented in the rest of the book and beyond.

Chapter 1

Models of Human Emotion and Artificial Emotional Intelligence



Benjamin Wortman

Abstract This chapter bridges the gap between emotion models popular in psychology and their use in affective computing tasks. Emotion modeling has a long and varied history with several competing schools of thought. Here, through a survey of existing literature, we cover existing emotion models popular in psychology, highlighting the strengths and weaknesses of these different approaches in regard to computational tasks involving human emotion.

1.1 Introduction

As an interdisciplinary field, affective computing sits at the corner of psychology and computing. This field has seen exponential growth in recent years¹ in tandem with the rise of Deep Learning and the creation of large scale datasets for training [10, 22, 33, 40, 51]. Although historically the focus has been on developing algorithms and techniques to help machines recognize and respond to human emotion, these technologies are underpinned by models in psychology. This is important to understand when transitioning to real-world settings as ultimately a machine-learning model is limited by the usefulness of the underlying emotion model. For example, Ekman's universal basic emotions [24] has long been the dominant model used for comparison in this field. In a lab setting, this has several benefits such as its ease of use and cross-cultural relevance. However, given it only consists of 7 emotions, this narrow coverage makes it difficult to develop real-world applications with high enough fidelity to accurately describe the wide variety of emotional states that humans present.

¹ Over 14,000 emotion recognition papers since 2010 according to IEEE Xplore.

B. Wortman (✉)

College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA

In many ways, identifying the correct human emotion model to use for affective computing tasks is a difficult challenge, especially given the differing philosophies between psychologists and computer scientists on what makes a theory of emotion useful. In psychology, the emphasis is on the correctness of the emotion model whereas in computer science the utility of the model is the priority. To help alleviate this, we provide a survey of existing literature to bridge the gap between models popular in psychology and their related computational tasks involving human emotion.

The rest of this chapter is outlined as follows. In Sect. 1.2, we cover emotion models from a psychological perspective. This includes cultural considerations, the relationship between emotion and personality, and finally an overview of the three competing schools of thought in human emotion modeling. In Sect. 1.3 we provide a review of existing literature to show how emotion models have been used in various affective computing tasks as well as the strengths and weaknesses of leveraging different models of emotion.

1.2 Emotion Modeling: A Psychological Perspective

As far back as Darwin [20], researchers have considered the subjectivity and universality of human emotion. This is a challenging problem as human emotion is highly complex and subjective. In fact, even defining what constitutes an emotion is still highly debated [30, 75, 82]. Although there has been no consensus on a scientific definition, in general, emotions can be thought of as a set of automatic, physiological, and psychological behaviors which allow an individual to address personally relevant situations [75]. Similar to heuristics, emotions are a sort of mental shortcut that allows individuals to automatically respond to stimuli without having to think about it. It has been widely accepted that emotions have been naturally selected over time since at one point in our evolutionary past they were effective in increasing the reproductive fitness of our ancestors [47]. For example,

Fig. 1.1 Emotions are thought to be evolutionarily derived. For example, a gazelle without any fear of predators would be unlikely to pass its genes down to the next generation (Image from Google search)



an animal afraid of potential predators (Fig. 1.1) is much more likely to survive and pass its genes down to the next generation. Likewise, similar evolutionary pressures are thought to be responsible for more complex emotions like love evolving in early mammals between mothers and their young [53], or shame evolving in our social ancestors to discourage behavior harmful to the survival of the group [88].

In the following section, we dive deeper into this topic by first providing a disambiguation between emotion and other affective phenomena. Then we discuss the three competing schools of thought for developing useful models of emotion. Finally, we cover relevant cultural considerations and discuss recent research into adjusting the definition of human emotion to account for physiological phenomena such as hunger, thirst, and sleep.

1.2.1 Distinguishing Emotion from Other Affective Phenomena

Although emotion is often used interchangeably with mood, feelings, or personality in everyday language, these affective states are actually quite different and can be discriminated from each other across using features [82]. In his taxonomy, Scherer uses event focus, appraisal, synchronization, rapidity of change, behavioral impact, intensity, and duration. Here event focus describes how whether the affective state is grounded by a specific event (internal or external) or whether it is simply free-floating, existing as a semi-permanent feature of the individual. The appraisal feature is derived from the component process model for emotion as discussed in Sect. 1.2.2.3 and is broken down into two categories: intrinsic and transactional. Intrinsic appraisal describes a person's preferences regarding the object or event being evaluated independent of their current needs (e.g. having a general preference for sweet foods). Conversely, transactional appraisal relates to how events with respect to how they satisfy the immediate needs of the appraiser such as reducing stress or anxiety [42]. Considering the adaptive roles emotion has on behavioral responses, synchronization refers to how mobilized the automatic response of the body's subsystems are to the triggering event or situation [81]. Similarly, behavioral impact describes how disruptive the affective state is to existing behavioral patterns. In Table 1.1 based on the work of Scherer [82], we can see how seven different affective phenomena relate across each of these dimensions. Descriptions for each of the states are as follows:

- **Preferences:** Judgements a person makes in regard to being attracted to or avoiding a certain stimulus. This is a relatively stable, low-intensity state and has little effect on behavior outside of liking or disliking the stimulus.
- **Attitudes:** A relatively enduring state with high intrinsic appraisal consisting of three pieces. First, a cognitive component consists of a belief about the stimulus or object. Second, an affective component describes whether the person has a

Table 1.1 Affective phenomena [82]

Affective phenomena	Features	Event focus	Intrinsic appraisal	Transactional appraisal	Synchronization	Rapidity of change	Behavioral impact	Intensity	Duration
Preferences	-	++	o	-	-	-	o	-	o
Attitudes	-	-	-	-	-	-	-	o	+
Moods	-	o	-	-	o	-	+	o	+
Personality traits	-	-	-	-	-	-	-	-	++
Interpersonal stances	+	-	-	-	-	++	+	o	o
Emotions (Utilitarian)	++	o	++	++	++	++	+	-	-
Emotions (Aesthetic)	+	++	-	o,+	-	-	-	o	-

Key: (-) Very low, (-) Low, (o) Medium, (+) High, (++) Very high

positive or negative feeling towards the stimulus. Finally, a behavioral component describing the action tendency with respect to the stimulus [11].

- **Moods:** Generally lacking a clear cause, moods are typically less intense than emotions, but have a longer duration possibly lasting several days at a time. Examples of moods include content, crabby, cheerful, and depressed.
- **Personality Traits:** Natural tendencies of a person towards certain affective dispositions. Emotion is thought to be a subsystem of this affective phenomenon since personality traits are considered stable and generalize across situations and stimuli [75]. A popular example of personality traits includes the Five-Factor Model which is a widely used taxonomy consisting of five semi-independent dimensions: neuroticism, extraversion, agreeableness, openness, and conscientiousness [34].
- **Interpersonal Stances:** This describes the affective style used during an interpersonal exchange in a given situation. Examples of this include being gregarious, polite, cold, distant, etc. Often triggered when engaging in social interaction, this state is influenced not just by the person's current attitude but also their strategic motivations for the interaction [82].
- **Emotions (Utilitarian):** Utilitarian emotions are what people typically would think of when describing emotion (e.g. happy, sad, angry, surprised). Being evolutionarily derived these typically serve some utilitarian function and as such have a high impact on a person's behavioral and bodily response to a stimulus.
- **Emotions (Aesthetic):** Unlike utilitarian emotions, aesthetic emotions are not adapted to fulfill any immediate need. Rather they are the product of an appreciation of the beauty of nature or artistic experiences [60, 84]. As such their effect on the body's subsystems is much less pronounced. Examples of aesthetic emotions include wonder, awe, bliss, admiration, or ecstasy.

1.2.2 *Three Competing Theories*

With a working definition of what constitutes an emotion, we can now begin to describe different models developed in psychology to describe human emotion. Generally, there are three competing theories: basic emotions, continuous models, and componential models. In the following sections, we provide an overview of each while highlighting their strengths and weaknesses.

1.2.2.1 Basic Emotion Theory

Basic emotion theory suggests that humans evolved a set of discrete, independent emotions which when triggered produce a physiological response or action tendency. From these basic emotions, all other human emotions can be derived. As shown in Table 1.2, these basic emotions are often used as categorical labels in affective computing datasets. More specifically, Paul Ekman's research into basic

Table 1.2 Recent emotion recognition datasets

Dataset	Labeled samples	Categorical emotions	Continuous emotions	Year
BoLD [51]	20k	26 ^a	VAD	2020
DFEW [33]	16k	7 ^b	–	2020
GoEmotions [22]	58k	28 ^b	–	2020
MOSEI [6]	23k	6 ^a	Sentiment	2018
OMG-Emotion [10]	0.6k	7 ^b	VA	2018
Aff-Wild [37]	0.3k	–	VA	2018
EMOTIC [40]	34k	26 ^a	VAD	2017
EmoReact [65]	1.1k	16 ^b	V	2016

^a Contains a subset of Ekman’s basic emotions

^b Ekman’s basic + neutral

Continuous Emotion Key: (V)alence, (A)rousal, (D)ominance

universal emotions serves as the foundation for most annotation schemes currently used [6, 10, 33, 40, 51, 65]. His original research identified six emotions universally recognizable by their facial expression [24]. They are fear, anger, joy, sadness, disgust, and surprise. However, several studies [9, 23, 45] suggest facial expressions alone cannot differentiate emotions. Since it has been demonstrated that body language cues are also universal across cultures [68], there may exist a subset of emotions that are universal for body language while being indistinguishable in facial expressions alone [16]. Although not shown to be cross-cultural, analysis by Cowen et al. on perceived emotions from vocalization [19], facial expressions [18], and perceived emotion from video [17] suggests not six but more than 24 emotion categories are required to adequately map the space. However, this was limited in that the label space was predetermined by the researchers. In attempting to develop an emotion model for text classification, Demszky et al. expanded upon Cowen’s work by using user-submitted labels to augment their emotion model. These labels were then pruned and refined to generate a more annotator-friendly list of 27 emotions and a neutral category [22].

Although Ekman’s basic emotions are the most commonly used in affective computing, other models do exist. In taking an evolutionary-inspired approach, Plutchik proposes an alternative to Ekman’s model which consists of eight primary affective states arranged to form a wheel of emotion [71]. Each of these affective states has varying degrees of intensity and when combined forms more complex human emotions. Although a useful tool, this model is criticized as being too simplistic and hasn’t been shown to have a strong empirical foundation [86]. Compared with Plutchik’s palette theory, Jaak Panksepp took a biological approach to understanding emotion. His work pioneered the field of affective neuroscience which works to map specific regions of the brain to emotional experience [46, 67, 70]. In his original work, he describes seven affective systems common across mammalian brains which control specific types of behaviors and generate distinct

emotional states [66]. He describes these structures as the “core-SELF”². Despite its neurological underpinnings this hasn’t been widely used in the affective computing community.

1.2.2.2 Continuous Dimensions of Emotion

In addition to discrete labels, work has been done in defining continuous dimensions to measure a person’s affective state. As shown in Table 1.2, annotations along continuous dimensions are often used together with basic emotions. In the simplest case, this can just be labeling a sample based on how positive or negative it is. This is usually described as the sentiment, pleasure, or valence of the sample. Expanding beyond one dimension, the Circumplex of Affect by Russel considers arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant) as the two fundamental dimensions which together provide a mapping for the discrete emotions [76]. There is strong support for the two-dimensional approach of the Affective Circumplex. These two dimensions appear across a wide range of studies [1, 77, 90]. A mapping of the Affective Circumplex based on video annotations from the BoLD dataset can be seen in Fig. 1.2. Similar to Panksepp’s mapping of discrete emotions, there has

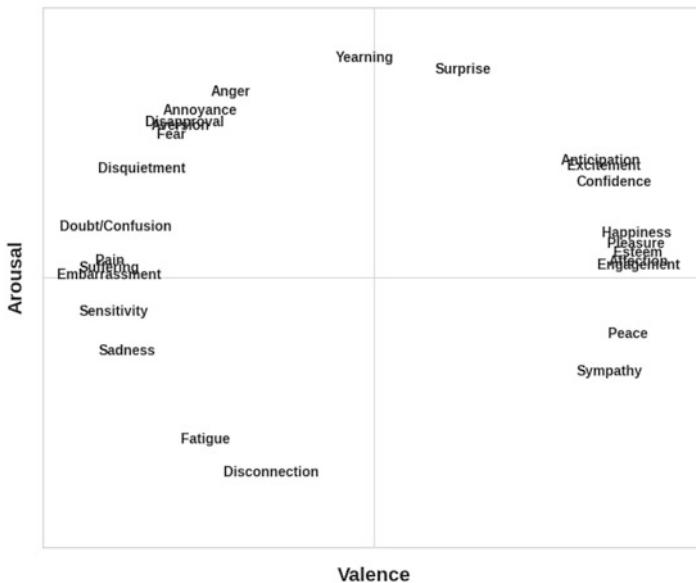


Fig. 1.2 The affective circumplex based on 26,378 annotations from the BoLD dataset [51]

² These include SEEKING (expectancy), FEAR (anxiety), RAGE (anger), LUST (sexual excitement), CARE (nurturing), PANIC/GRIEF (sadness), and PLAY (social joy).

also been a considerable amount of work mapping valence and arousal to processes in the human brain [4, 35, 72, 78].

For three dimensions, another popular model comes from the researchers Mehrabian et al. who described the emotion space across pleasure-displeasure, arousal-nonarousal, and dominance-submissiveness (PAD³) [59]. This mirrors earlier work by Osgood et al., who considered the closely related concept of control instead of dominance [15]. Here, control can be thought of in terms of both the feelings of power or weakness in addition to interpersonal dominance or submission. With regards to the PAD model, other proposed dimensions include anticipation-expectation, anxiety-confidence, boredom-fascination, frustration-euphoria, terror-enchantment, and intensity (how far the person is from a state of pure, cool rationality) [26, 38, 58].

1.2.2.3 Componential Theories of Emotion

In contrast with discrete basic emotions and the previously described continuous models, there has also been some work in developing componential models derived from the appraisal theory of emotion [26, 28, 55, 85]. This is the dominant theory for describing how emotions are generated [64]. Under this framework, emotion is not a state but a dynamic process thought to result from a person's repeated evaluation (appraisal) of their circumstances [5, 43]. This process is broken down into several components including appraisal, action tendency, bodily reaction, expression, and feeling [85]. An example of this process is shown in Fig. 1.3. To generate emotions a person first evaluates the scenario they are in. Their central nervous system then prepares an action (e.g. a fight or flight response). Bodily symptoms present themselves such as a change in heartbeat, shivers, or blushing. Similarly, there is a change in motor expression such as changes in speech, changes in body language, or facial expressions [83]. Finally, these changes are manifested as feeling which can be described by their intensity, duration, valence, arousal, and tension [85].

This has an advantage over descriptive models leveraging basic emotions or continuous dimensions by providing an explanation for why an emotion presents itself. However, since these componential models rely heavily on subjective expe-



Fig. 1.3 An example of the processes within the componential model of emotion according to Scherer [85]

³ Valence and pleasure are often used interchangeably so this is also sometimes referred to as VAD for valence, arousal, and dominance.

rience [26, 80], outside of lab-constrained experiments [62, 63] they have not been widely adopted for use in affective computing.

1.2.3 Cultural Considerations

When discussing affective computing it is also important to consider that emotional perception and expression are heavily influenced by language and culture. For example, despite Ekman and Friesen identifying several facial expressions that are universally recognized as belonging to a certain emotion [25], they also identified display rules specific to a person's culture that governed their expression of these emotions. A study by Malatesta et al. in 1984 [54] appears to confirm this finding evidence to suggest these display rules are learned in infancy. As an example of how display rules may differ, consider a simple smile (Fig. 1.4). Although in the United States smiling might make others perceive you as happier and more attractive [74], even in Norway or Poland where the cultural distance is not as extreme as 'the East' vs. 'the West', smiling is perceived as a sign of stupidity. In fact, in Russian, there is actually a saying translated as "*smiling with no reason is a sign of stupidity*" [41]. In addition to the meaning behind these displays, the attribution of emotional expression in individualist vs. collectivist cultures has also been shown to differ. In collectivist cultures, people are more likely to view others' emotional expressions as being directly relevant to the status of their relationship with that person. This differs from individualistic cultures where another person's emotions are not necessarily directed at them [61]. For example, in a collectivist culture, a person who sees their friend with an angry expression might perceive their friend as being angry at them.

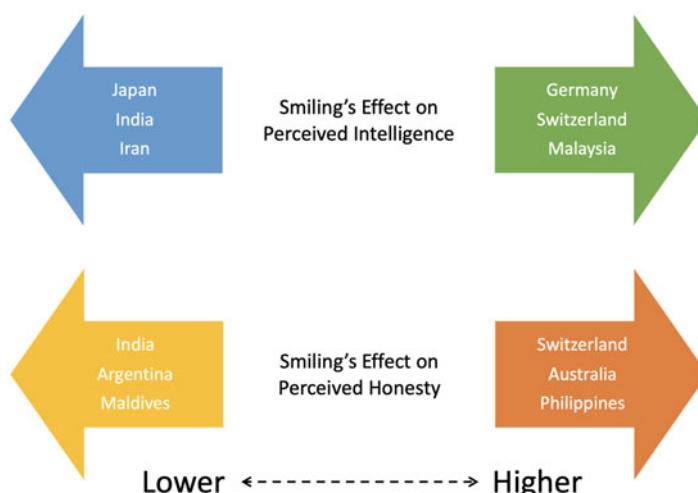


Fig. 1.4 Smiling can be interpreted differently depending on what country you are in [41]

Whereas in an individualistic culture, people would not necessarily consider their friend's anger as being related to themselves.

Hand in hand with culture, language similarly plays a key role in the perception of emotion. Language has evolved to be the principal means of human communication and has been shown to influence our perception of the world [8, 48]. The simplest example of how language influences our perception of emotion is the fact that there are numerous examples of emotional concepts having no direct translation across languages [50]. For example, the German word Schadenfreude is a popular example of an emotionally charged word with no English equivalent. In addition to the limitations in vocabulary, the emotional impact of words also varies by geographic region [7, 32]. These differences across cultures make generating universal cross-cultural applications in affective computing difficult. Even for universally recognized emotions like happiness and sadness, the differences across cultures and languages make annotations difficult since annotators from different regions may have a fundamental disagreement on the meaning of the labels. This suggests research or development toward emotion-related tasks would benefit from a more localized approach. This is especially true given the fact that local annotators are better at recognizing emotional expressions within their own culture [31].

1.2.4 *Physiological Considerations*

In addition to traditionally understood emotions, recent work has formalized the relationship between emotional states and other physiological processes such as sleep [57], hunger [89], thirst [52], and sex drive [14]. Emotions and physiological motivations are closely related as both are considered evolutionarily-derived phenomena for promoting survival in natural environments [47]. However, unlike emotions which are fleeting and context-dependent, these physiological processes are permanent throughout a person's life. In the case of thirst, hunger, and sleep; these motivations help maintain a person's homeostasis [21]. For example, in Maslow's hierarchy of needs, these motivations form the base of the pyramid. When these needs are not met, higher-order needs such as socialization, self-esteem, or self-actualization become subordinate until those basic needs are met [56]. In the context of affective computing, given the relationship between these motivations and emotions, it may be helpful to consider these when developing emotion-related technologies or carrying out emotion-related research.

1.3 Applications in Computing

We can break down research in affective computing that is related to emotion models into four general categories. The first broadly aligns with traditional supervised machine learning techniques. Here emotion is used as a labeled ground truth target

for training machine learning models to classify human emotional expressions. The second category is work developing methods to synthesize emotional responses from machine intelligence to aid in more natural human-machine interaction. The third uses machine learning techniques to identify statistical trends and develop new statistically grounded models of human emotion. Finally, we explore exciting opportunities in emergent behavior from large transformer and diffusion models. We highlight each of these in the following section.

1.3.1 *Emotion as a Target for Training*

Using emotion labels as a target for training is the most straightforward application of emotion models in affective computing. Given a labeled dataset, researchers use supervised or semi-supervised techniques to train their models to recognize human affective states. Targets for this task typically include continuous dimensions such as the Affective Circumplex [76] or discrete states such as Ekman’s basic emotions [69]. This can leverage any number of modalities including images [40], video [51], text [22], and audio [49]. A summary of recent emotion recognition datasets is provided in Table 1.2.

When selecting emotion labels for these tasks, it is worth noting that in practice, there is a trade-off between the size of the emotion model and the amount of information the annotators will give for any sample. As emotion models become more complex and include more abstract concepts, the agreement between annotators decreases substantially. This reduction can be seen in the levels of inter-annotator agreement across several large-scale datasets that reported this information. For example, despite several quality-control measures implemented during data collection and the post-processing done to filter unreliable annotators in the BoLD dataset [51], the average Fleiss’ Kappa [29] across emotion categories is $\kappa = 0.173$ [51]. Intuitively less complex emotions like “Happiness” have higher levels of agreement comparable to objective tasks performed at the time of data collection like determining age or ethnicity. More abstract concepts, such as “Yearning” and “Sensitivity,” had almost no agreement among participants. This result mirrors those from the EMOTIC [40] and GoEmotions [22] datasets. This comparison is important since as the size of the emotion model increases the types of emotion concepts included are bound to become more abstract. Not only do additional components have diminishing returns in terms of the information they provide, but since emotion is subjective, they also suffer an agreement penalty during the annotation process further reducing their effectiveness. One method to help mitigate this effect would be to choose more concrete emotion concepts as the foundation for emotion models.

1.3.2 Mimicking Human Emotion

One of the primary goals in affective computing is to foster more natural human-computer interaction. However, communication is a two-way street, and recognizing emotions is only the first part of this. This is why some researchers have turned toward developing machines that can mimic human expression. This takes on a variety of forms and has potential applications in animation and social robotics.

For emotion models used in this type of research, some researchers [44, 79, 87] have found value in vectored approaches using Russel's affective circumplex [76]. Since this is a vector model, in theory, this is easy to work with since simulating different emotions is as easy as adjusting the input vector provided to the machine. Compared with basic emotions, this continuous model makes it easier to represent a wider range of emotional expressions. The actual synthesis of these expressions has benefited from annotation systems for describing the movements of the human body. In facial expression generation, the Facial Action Coding System (FACS) is commonly referenced since it provides a complete mapping of the human face in the form of action units (AU) [69]. These AUs describe the movements of individual muscles and can be used to represent distinct emotions [27]. For example, AUs related to turning down the edges of your mouth while raising your inner eyebrow and lowering the outer eyebrow relate to the facial expression associated with frowning or sadness. Although a FACS-like system is not currently available for the full human body, techniques leveraging Laban Movement Analysis (LMA) in a similar manner have shown some success in synthesizing emotive gestures [13]. Laban Movement Analysis was first developed for use in dance and comprises of four components: Body, Space, Effort, and Shape. Together these provide a comprehensive system for annotating and characterizing the quality of movement. In discussing body emotion recognition, it is worth mentioning related work in gesture synthesis which has had quite a bit of success disentangling and transferring gestures styles from speaker to speaker [2]. This benefits from large datasets in somewhat restricted settings (e.g. talk shows) [3] so it is much easier for the model to learn discriminant features. Since gestures naturally correspond to speech, then using a similar methodology of emotive gesture synthesis from emotionally charged speech may be possible given a large enough training set.

1.3.3 Statistical Techniques for Developing Emotion Models

Unlike the previous two sections where predetermined emotion models were used to proved researchers with a set of labels for their work, in this section we cover research into the actual development of these emotion label sets. In general research in this area can be broken down into two categories: fixed and open-ended.

Fixed studies start with a predetermined set of emotion labels and then set out to statistically prove independence or the ability to discriminate between emotions.

This typically takes on the form of having annotators label samples and then performing statistical analysis on their responses. These make up the majority of literature in the emotion model space. Examples of this include Cowen et al.'s studies into perceived emotions from vocalization [19], facial expressions [18], and video [17]. In each of these, the test subjects are given a piece of media and asked to identify the emotions associated with it from a predetermined list, and then tests are run to see if these predictions occur at a rate above random chance. In some instances these are augmented with free response studies, however, since the original data collection is associated with a fixed group of emotional states⁴ these are still restricted in the emotions that can potentially be elicited. In a meta-analysis of similar studies, Keltner et al. identified 24 emotions that are discriminated from each other at an above random chance rate in at least one modality [36]. However, the authors are quick to point out that in their meta-analysis that there is a possibility for additional emotions in multi-modal contexts.

In contrast with fixed labeled studies, open-ended studies are not limited by any predetermined label set and rely more on clustering and dimensionality reduction techniques than annotators to identify latent dimensions of emotion. An example of this type of study would be the development of HICEM, a HIgh Coverage Emotion Model for use in affective computing [91]. Here natural language word embeddings are used to encode the semantic meaning of 1720 emotion-related concepts. Then using a combination of dimensionality reduction and hierarchical clustering, a list of 15 components is generated which provides the highest coverage across the entire emotion concepts list (Fig. 1.5).

1.3.4 *Emergent AEI*

The incredible performance of large-scale transformer-based models is one of the most exciting developments in AI in recent years. Although these models are not trained specifically for emotion-related tasks, they have a strong ability to generalize due to being trained on massive, internet-collected datasets. For example, despite not explicitly being trained for emotion generation, diffusion models trained on text and image pairs like Dalle-2 [73] can produce photorealistic images of people expressing basic emotions as shown in Fig. 1.6.

On the NLP front, GPT-4 likewise has shown an incredible understanding of human emotion. Using 40 tasks designed to test theory of mind capabilities, GPT-4 was able to successfully solve 95% of them giving it comparable performance to a 9-year-old child [39]. This metacognition allows the model to provide plausible responses when queried about others' motivations and emotional states [12].

⁴ For example in Cowen et al.'s study on self-reported emotions from video, the videos were gathered from prompts based on 34 predetermined emotion categories.

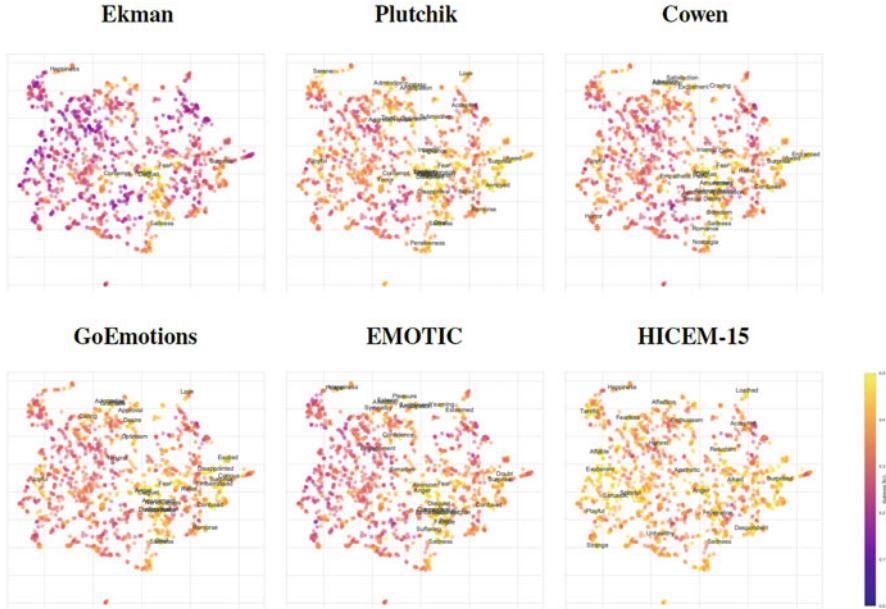


Fig. 1.5 The maximum log cosine similarity (with a ceiling of 0.5) between the word vectors of 1720 emotion concepts and the contents of the model from the Wortman and Wang’s HICEM paper [91]. A higher cosine similarity (yellow) means the contents of the model have a similar semantic meaning to the given concept. Let k denote the number of labels or components in an emotion model. From top left going clockwise Ekman’s basic emotions [24] ($k = 7$), Plutchik’s wheel of emotion [71] ($k = 32$), Cowen’s emotions identified in video [17] ($k = 27$), HICEM-15 [91], the EMOTIC [40] dataset annotation scheme ($k = 27$), and the annotation categories for the GoEmotions Dataset [22] ($k = 28$)

Both Dalle-2 and GPT4 have shown the power of large transformer-based models and their ability to implicitly encode emotion information. Future applications can leverage this to provide more natural and customizable interactions. Still, there are open questions about ethics, alignment, and the psychological impact these models have on human users.

1.4 Conclusion

In many ways, identifying the correct human emotion model to use for affective computing tasks is a difficult challenge. Between cultural considerations and the subjectivity of emotional experience, universal solutions remain elusive. However, advances in both implicit and explicit emotional understanding continue to advance the dream of seamless human-machine interaction.



Fig. 1.6 In the above images generated by Dalle-2 [73], the model is able to effectively synthesize representative expressions for basic emotions. These images were all generated using the prompt “Portrait of a <emotion> man”

Acknowledgments The work was funded in part by a generous gift from Amazon to the author’s dissertation advisor Professor James Z. Wang. The author also acknowledges the advice and constructive comments from James Wang, Reginald Adams, Jr., and Tal Shafir.

References

1. Abelson, R.P., Sermat, V.: Multidimensional scaling of facial expressions. *J. Exp. Psychol.* **63**(6), 546–554 (1962)
2. Ahuja, C., Lee, D.W., Ishii, R., Morency, L.P.: No gestures left behind: learning relationships between spoken language and freeform gestures. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp. 1884–1895 (2020)
3. Ahuja, C., Lee, D.W., Nakano, Y.I., Morency, L.P.: Style transfer for co-speech gesture animation: a multi-speaker conditional-mixture approach (2020). CoRR abs/2007.12553 ArXiv: 2007.12553
4. Anderson, A., Christoff, K., Stappen, I., Panitz, D., Ghahremani, D.G., Glover, G., Gabrieli, J., Sobel, N.: Dissociated neural representations of intensity and valence in human olfaction. *Nat. Neurosci.* **6**(2), 196–202 (2003)
5. Arnold, M.B.: *Emotion and Personality Psychological Aspects*, vol. 1. Columbia University Press, New York (1960)

6. Bagher, Z.A., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Long Papers, vol. 1, pp. 2236–2246. Association for Computational Linguistics, Melbourne (2018)
7. Bann, E.Y., Bryson, J.J.: Measuring cultural relativity of emotional valence and arousal using semantic clustering and twitter (2013). CoRR abs/1304.7507. *_eprint: 1304.7507*
8. Barrett, L.F., Lindquist, K.A., Gendron, M.: Language as context for the perception of emotion. *Trends Cognitive Sci.* **11**(8), 327–332 (2007)
9. Barrett, L.F., Mesquita, B., Gendron, M.: Context in emotion perception. *Curr. Dir. Psychol. Sci.* **20**(5), 286–290 (2011)
10. Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., Wermter, S.: The OMG-emotion behavior dataset. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE, Rio de Janeiro (2018)
11. Breckler, S.J.: Empirical validation of affect, behavior, and cognition as distinct components of attitude. *J. Pers. Soc. Psychol.* **47**(6), 1191–1205 (1984)
12. Bubeck, S.A., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: early experiments with GPT-4 (2023). ArXiv:2303.12712 [cs]
13. Burton, S.J., Samadani, A.A., Gorbet, R., Kulic, D.: Laban movement analysis and affective movement generation for robots and other near-living creatures. In: Dance Notations and Robot Motion, pp. 25–48. Springer, Berlin (2016)
14. Burunat, E.: Love is a physiological motivation (like hunger, thirst, sleep or sex). *Med. Hypotheses* **129**, 109225 (2019)
15. Carroll, J.B., Osgood, C.E., May, W.H., Miron, M.S.: Cross-cultural universals of affective meaning. *Am. J. Psychol.* **89**(1), 172 (1976)
16. Cordaro, D.T., Sun, R., Kamble, S., Hodder, N., Monroy, M., Cowen, A., Bai, Y., Keltner, D.: The recognition of 18 facial-bodily expressions across nine cultures. *Emotion* **20**(7), 1292–1300 (2020)
17. Cowen, A.S., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci.* **114**(38), E7900–E7909 (2017)
18. Cowen, A.S., Keltner, D.: What the face displays: mapping 28 emotions conveyed by naturalistic expression. *Am. Psychol.* **75**(3), 349–364 (2020)
19. Cowen, A.S., Elfenbein, H.A., Laukka, P., Keltner, D.: Mapping 24 emotions conveyed by brief human vocalization. *Am. Psychol.* **74**(6), 698–712 (2019)
20. Darwin, C.: *The Expression of the Emotions in Man and Animals*. University of Chicago Press, Chicago (2015)
21. Davies, K.J.: Adaptive homeostasis. *Mol. Aspects Med.* **49**, 1–7 (2016)
22. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: GoEmotions: a dataset of fine-grained emotions. In: 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4040–4054 (2020)
23. Ekman, P.: Facial expression and emotion. *Am. Psychol.* **48**(4), 384–392 (1993)
24. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **17**(2), 124. American Psychological Association, Washington (1971)
25. Ekman, P., Friesen, W.V.: *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*, 2. [pr.] edn. Prentice-Hall, Englewood Cliffs (1975). OCLC: 247971765
26. Fontaine, J.R., Scherer, K.R., Roesch, E.B., Ellsworth, P.C.: The world of emotions is not two-dimensional. *Psychol. Sci.* **18**(12), 1050–1057 (2007)
27. Friesen, W., Ekman, P.: EMFACS-7: Emotional Facial Action Coding System. Unpublished manuscript, vol. 2, p. 1. University of California at San Francisco (1983)
28. Grandjean, D., Sander, D., Scherer, K.R.: Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Conscious. Cogn.* **17**(2), 484–495 (2008)
29. Gwet, K.L.: *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*, 4th edn. Advances Analytics, LLC, Gaithersburg (2014)

30. Izard, C.E.: Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspect. Psychol. Sci.* **2**(3), 260–280 (2007)
31. Jack, R.E., Caldara, R., Schyns, P.G.: Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *J. Exp. Psychol. General* **141**(1), 19–25 (2012)
32. Jackson, J.C., Watts, J., Henry, T.R., List, J.M., Forkel, R., Mucha, P.J., Greenhill, S.J., Gray, R.D., Lindquist, K.A.: Emotion semantics show both cultural variation and universal structure. *Science* **366**(6472), 1517–1522 (2019)
33. Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J.: DFEW: A large-scale database for recognizing dynamic facial expressions in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2881–2889 (2020)
34. John, O.P., Robins, R.W., Pervin, L.A.: *Handbook of Personality: Theory and Research*. Guilford Press, New York (2010)
35. Jones, B.E.: Arousal systems. *Front. Biosci.* **8**(6), s438–s451 (2003)
36. Keltner, D., Sauter, D., Tracy, J., Cowen, A.: Emotional expression: advances in basic emotion theory. *J. Nonverbal Behav.* **43**(2), 133–160 (2019)
37. Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B.A.W., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond (2018). CoRR abs/1804.10938. ArXiv: 1804.10938
38. Kort, B., Reilly, R., Picard, R.: An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In: Proceedings IEEE International Conference on Advanced Learning Technologies, pp. 43–46 (2001)
39. Kosinski, M.: Theory of Mind May Have Spontaneously Emerged in Large Language Models (2023). ArXiv:2302.02083 [cs]
40. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: EMOTIC: emotions in context dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2309–2317. IEEE, Honolulu (2017)
41. Krys, K., Melanie Vauclair, C., Capaldi, C.A., Lun, V.M.C., Bond, M.H., Domínguez-Espínosa, A., Torres, C., Lipp, O.V., Manickam, L.S.S., Xing, C., et al.: Be careful where you smile: culture shapes judgments of intelligence and honesty of smiling individuals. *J. Nonverbal Behav.* **40**, 101–116 (2016)
42. Lazarus, R.S.: Cognition and motivation in emotion. *Am. Psychol.* **46**(4), 352. American Psychological Association, Washington (1991)
43. Lazarus, R.S.: *Psychological Stress and the Coping Process*. McGraw-Hill, New York (1966)
44. Lazzeri, N., Mazzei, D., Cominelli, L., Cisternino, A., De Rossi, D.: Designing the mind of a social robot. *Appl. Sci.* **8**(2), 302 (2018)
45. Le Mau, T., Hoemann, K., Lyons, S.H., Fugate, J.M.B., Brown, E.N., Gendron, M., Barrett, L.F.: Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs. *Nat. Commun.* **12**(1), 5037 (2021)
46. LeDoux, J.E.: Emotion circuits in the brain. *Ann. Rev. Neurosci.* **23**(1), 155–184 (2000)
47. LeDoux, J.E.: Chapter 21 - evolution of human emotion: a view through fear. In: M.A. Hofman, D. Falk (eds.) *Progress in Brain Research, Evolution of the Primate Brain*, vol. 195, pp. 431–442. Elsevier, Amsterdam (2012)
48. Lindquist, K.A., Gendron, M.: What's in a word? language constructs emotion perception. *Emot. Rev.* **5**(1), 66–71 (2013)
49. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS One* **13**(5), e0196391 (2018)
50. Lomas, T.: Towards a cross-cultural lexical map of wellbeing. *J. Posit. Psychol.* 1–18 (2020)
51. Luo, Y., Ye, J., Adams, R.B., Jr., Li, J., Newman, M.G., Wang, J.Z.: ARBEE: towards automated recognition of bodily expression of emotion in the wild. *Int. J. Comput. Vision* **128**(1), 1–25 (2020)
52. MacCormack, J.K., Lindquist, K.A.: Feeling hangry? When hunger is conceptualized as emotion. *Emotion* **19**(2), 301–319 (2019)

53. MacDonald, K., Patch, E.A., Figueiredo, A.J.A.: Love, trust, and evolution: nurturance/love and trust as two independent attachment systems underlying intimate relationships. *Psychology* **07**(02), 238–253 (2016)
54. Malatesta, C.Z., Haviland, J.M.: Learning display rules: the socialization of emotion expression in infancy. *Child Dev.* **53**(4), 991 (1982)
55. Marinier, R.P., Laird, J.E., Lewis, R.L.: A computational unification of cognitive behavior and emotion. *Cogn. Syst. Res.* **10**(1), 48–69 (2009)
56. Maslow, A.H.: *A Dynamic Theory of Human Motivation*, pp. 26–47. Howard Allen Publishers, Cleveland (1958)
57. McGlinchey, E.L., Talbot, L.S., Chang, K.h., Kaplan, K.A., Dahl, R.E., Harvey, A.G.: The effect of sleep deprivation on vocal expression of emotion in adolescents and adults. *Sleep* **34**(9), 1233–1241 (2011)
58. McKeown, G., Valstar, M.F., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: 2010 IEEE International Conference on Multimedia and Expo, pp. 1079–1084 (2010)
59. Mehrabian, A.: Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in Temperament. *Curr. Psychol.* **14**(4), 261–292 (1996)
60. Menninghaus, W., Wagner, V., Wassiliwizky, E., Schindler, I., Hanich, J., Jacobsen, T., Koelsch, S.: What are aesthetic emotions? *Psychol. Rev.* **126**(2), 171–195 (2019)
61. Mesquita, B.: Emotions in collectivist and individualist contexts. *J. Pers. Soc. Psychol.* **80**(1), 68–74 (2001)
62. Meuleman, B., Rudrauf, D.: Induction and profiling of strong multi-componential emotions in virtual reality. *IEEE Trans. Affect. Comput.* **12**(1), 189–202 (2021)
63. Mohammadi, G., Vuilleumier, P.: A multi-componential approach to emotion recognition and the effect of personality. *IEEE Trans. Affect. Comput.* 1–1 (2020)
64. Moors, A., Ellsworth, P.C., Scherer, K.R., Frijda, N.H.: Appraisal theories of emotion: state of the art and future development. *Emot. Rev.* **5**, 119–124 (2013)
65. Nojavanaghari, B., Baltrušaitis, T., Hughes, C.E., Morency, L.P.: EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 137–144 (2016)
66. Panksepp, J.: *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, Oxford (2004)
67. Panksepp, J., Biven, L.: *The Archaeology of Mind: Neuroevolutionary Origins of Human Emotions*. A Norton Professional Book, 1st edn. W. W Norton, New York (2012)
68. Parkinson, C., Walker, T.T., Memmi, S., Wheatley, T.: Emotions are understood from biological motion across remote cultures. *Emotion* **17**(3), 459–477 (2017)
69. Paul, E., Wallace, F.: *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto (1978)
70. Phan, K., Wager, T., Taylor, S.F., Liberzon, I.: Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage* **16**(2), 331–348 (2002)
71. Plutchik, R.: *The Emotions*, revised edn. University Press of America, Lanham (1991)
72. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **17**(03) (2005)
73. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022)
74. Reis, H.T., Wilson, I.M., Monestere, C., Bernstein, S., Clark, K., Seidl, E., Franco, M., Gioioso, E., Freeman, L., Radoane, K.: What is smiling is beautiful and good. *Eur. J. Soc. Psychol.* **20**(3), 259–267 (1990)
75. Reisenzein, R., Hildebrandt, A., Weber, H.: Personality and emotion. In: G. Matthews, P.J. Corr (eds.) *The Cambridge Handbook of Personality Psychology*, Cambridge Handbooks in Psychology, 2 edn., pp. 81–100. Cambridge University Press, Cambridge (2020)
76. Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161–1178 (1980)

77. Russell, J.A., Bullock, M.: Multidimensional scaling of emotional facial expressions: similarity from preschoolers to adults. *J. Pers. Soc. Psychol.* **48**(5), 1290–1298 (1985)
78. Sackheim, H.A.: Hemispheric asymmetry in the expression of positive and negative emotions: neurologic evidence. *Arch. Neurol.* **39**(4), 210 (1982)
79. Saldien, J., Goris, K., Vanderborght, B., Vanderfaeillie, J., Lefeber, D.: Expressing emotions with the social robot probo. *Int. J. Soc. Robot.* **2**(4), 377–389 (2010)
80. Sander, D., Grandjean, D., Scherer, K.R.: A systems approach to appraisal mechanisms in emotion. *Neural Netw.* **18**(4), 317–352 (2005)
81. Scherer, K.R.: Emotions as episodes of subsystem synchronization driven by nonlinear appraisal processes. In: M.D. Lewis, I. Granic (eds.) *Emotion, Development, and Self-Organization*, 1st edn., pp. 70–99. Cambridge University Press, Cambridge (2000)
82. Scherer, K.R.: What are emotions? And how can they be measured? *Soc. Sci. Inf.* **44**(4), 695–729 (2005)
83. Scherer, K.R., Fontaine, J.R.J.: The semantic structure of emotion words across languages is consistent with componential appraisal models of emotion. *Cogn. Emot.* **33**(4), 673–682 (2019)
84. Scherer, K., Zentner, M.: Music evoked emotions are different—more often aesthetic than utilitarian. *Behav. Brain Sci.* **31**(5), 595–596 (2008)
85. Scherer, K.R., Schorr, A., Johnstone, T. (eds.): *Appraisal Processes in Emotion: Theory, Methods, Research*. Series in Affective Science. Oxford University Press, Oxford/New York (2001)
86. Smith, H., Schneider, A.: Critiquing models of emotions. *Sociol. Methods Res.* **37**(4), 560–589 (2009)
87. Stock-Homburg, R.: Survey of emotions in human-robot interactions: perspectives from robotic psychology on 20 years of research. *Int. J. Soc. Robot.* **14**(2), 389–411 (2022)
88. Szycer, D., Tooby, J., Cosmides, L., Porat, R., Shalvi, S., Halperin, E.: Shame closely tracks the threat of devaluation by others, even across cultures. *Proc. Natl. Acad. Sci.* **113**(10), 2625–2630 (2016)
89. Verma, D., Wood, J., Lach, G., Herzog, H., Sperk, G., Tasan, R.: Hunger promotes fear extinction by activation of an amygdala microcircuit. *Neuropsychopharmacology* **41**(2), 431–439 (2016)
90. Watson, D., Wiese, D., Vaidya, J., Tellegen, A.: The two general activation systems of affect: structural findings, evolutionary considerations, and psychobiological evidence. *J. Pers. Soc. Psychol.* 820–838 (1999)
91. Wortman, B., Wang, J.Z.: HICEM: a high-coverage emotion model for artificial emotional intelligence. In: IEEE Transactions on Affective Computing (2022). <https://doi.org/10.1109/TAFFC.2023.3324902>

Chapter 2

A Concise Introduction to Machine Learning



Sitao Zhang

Abstract Machine learning techniques have gained significant prominence in diverse multidisciplinary research fields, encompassing computational modeling of aesthetics, emotion, and artistic style, among others. This chapter serves as a tutorial to provide researchers, who may have limited or no prior expertise in the machine learning community, with essential background knowledge. We commence by defining learning algorithms and subsequently explore fundamental concepts and principles underlying machine learning systems. Furthermore, we present a concise introduction to several typical and basic machine learning models. In order to cater to novice readers, particularly those without formal training in computing, machine learning, or statistical modeling, we deliberately refrain from extensive formula derivations. Given the pervasive utilization of machine learning in the modeling of aesthetics, emotion, and artistic style, grasping the fundamentals of machine learning becomes instrumental in comprehending recent research endeavors within these domains. By offering an overview, this chapter aims to spark interest and encourage readers to delve into more comprehensive and in-depth coverage of the fundamentals through relevant textbooks and recent research papers.

2.1 Learning Algorithms

A machine learning algorithm is an algorithm that is able to learn from data. But what do we mean by “learning”? Mitchell provides a succinct definition, “*A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.*” [8] Throughout this chapter, we will delve into the precise definitions of these concepts, while also introducing commonly employed tasks, performance measures, and experiences through intuitive examples. These

S. Zhang (✉)

College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA

e-mail: sitao.zhang@psu.edu

illustrative examples serve as a foundation that can be extended to a wide array of applications, including the computational modeling of aesthetics, emotion, and artistic style.

2.1.1 *The Task, T*

In general, tasks in machine learning are closely associated with the specific problems we aim to address. Formally, learning itself is not considered a task since it represents the process of acquiring the capability to perform a task. For instance, if our objective is to enable robots to create art, the task at hand could be painting. We may train robots to learn how to draw, or we can directly generate artwork using predefined rules. However, for complex tasks related to artistic style, defining a set of rules may be impractical.

The advantage of machine learning, often referred to as “learning from data”, lies in its ability to tackle challenges that are difficult for rule-based deterministic systems. Consequently, the task of machine learning is typically characterized by processing samples. A sample represents a collection of quantitatively measured features for an item or event that the machine learning system will analyze. Conventionally, we represent a sample in \mathbb{R}^n as a vector x , where each entry x_i corresponds to a distinct feature. For instance, the fundamental features of an image often comprise the pixel values, while higher-level features can encompass visual characteristics derived from the pixel values.

Machine learning can accomplish a wide variety of tasks, some common forms are as follows:

- **Classification:** The goal of a classification task is to classify the input into one of the predefined categories to which it belongs. Learning algorithms usually require to learn a function $f : \mathbb{R}^m \rightarrow \{1, \dots, k\}$. In some cases, a sample may also have multiple labels.
- **Regression:** The goal of a regression task is to predict a real value for a given input. Learning algorithms usually require learning a function $f : \mathbb{R}^m \rightarrow \mathbb{R}$. This task has many similarities with classification in general.
- **Generation:** Tasks of this nature involve generating new samples that bear a resemblance to the training data. This typically necessitates a learning algorithm that can effectively model the data distribution and generate samples accordingly. In some cases, we may also seek the ability of the model to generate a specific type of output for a given input.
- **Anomaly Detection:** Anomaly detection tasks primarily focus on identifying rare items, events, or observations that significantly deviate from the majority of the data and do not conform to a well-defined notion of normal behavior.
- **Clustering:** These tasks involve grouping a set of objects in such a way that objects in the same group are in some sense more similar to each other than to those in other groups.

The tasks mentioned here serve as typical examples to illustrate the capabilities of machine learning, rather than an exhaustive taxonomy of tasks. In fact, machine learning can be applied to almost any task as long as the problem can be reasonably formulated by the researcher.

2.1.2 *The Performance Measure, P*

In order to evaluate the capabilities of a machine learning algorithm, it is necessary to establish a quantifiable performance measure, which is typically specific to the task being performed (referred to as T).

For common classification tasks, accuracy is often employed as a performance measure to assess the model's effectiveness. Accuracy is defined as the proportion of samples for which the model produces the correct output. However, it is important to note that the choice of measurement heavily depends on the specific context and data points. Accuracy can be misleading when dealing with imbalanced datasets. To illustrate this, consider a dataset consisting of 95 negative samples and 5 positive samples. Classifying all samples as negative would yield an accuracy score of 0.95, but clearly, this is not an accurate or desirable model. To address this issue, alternative measures that are unaffected by class imbalances can be utilized. For instance, balanced accuracy normalizes true positive and true negative predictions based on the respective number of positive and negative samples. In certain medical-related tasks, recall (also known as sensitivity) is often of greater importance. This is because missing a single case of a critical disease can have severe consequences, and therefore, a certain misdiagnosis rate is acceptable. It is worth noting that many of these classification metrics rely on labels and predictions, and may not be applicable to other tasks such as generation or clustering, which require different evaluation metrics.

In general, the focus is often placed on the generalization capabilities of a machine learning algorithm, which refer to its ability to perform well on unseen data, thus determining its effectiveness in real-world applications. Typically, a separate test set is used to represent unseen data, which is distinct from the training set used to train the machine learning system.

Overall, selecting an appropriate performance measure may initially appear straightforward and objective, but it can often be a complex task. In many cases, it is necessary to tailor measures to the specific task at hand or employ multiple distinct metrics to systematically evaluate the model's performance across various aspects.

2.1.3 *The Experience, E*

Machine learning algorithms can be broadly classified into two categories based on the type of experience they gain during the learning process: unsupervised learning and supervised learning.

Unsupervised learning methods are typically employed when working with datasets that only contain features, without any associated labels or targets. The objective of unsupervised learning is to extract meaningful structural information from the dataset. Understanding the overall distribution of the data is often crucial for tasks such as denoising, synthesis, clustering, and more.

On the other hand, supervised learning algorithms operate on datasets that include both features and corresponding labels or targets. These labels are closely related to the specific task of interest. For example, the ImageNet dataset provides annotations indicating the category to which each image belongs, enabling the training of models for image classification, object detection, and even generation tasks.

Unsupervised learning involves learning the underlying distribution, denoted as $p(x)$, by observing a collection of samples from that distribution. In contrast, supervised learning generates predictions by observing pairs of samples (x, y) and often estimates the conditional distribution $p(y|x)$. The primary distinction lies in whether explicit supervision signals are present in the dataset.

It's important to note that the definitions of unsupervised learning and supervised learning are not rigidly defined. There exists a degree of overlap and conversion between these two forms in many cases. In recent years, other variants of learning paradigms have emerged. Semi-supervised learning deals with datasets where only some samples are labeled, while others are not. Self-supervised learning leverages augmented transformations applied to samples as a source of supervision signals. Domain adaptation focuses on scenarios where the target domain lacks labeled samples, but the source domain does. Additionally, not all machine learning algorithms are trained on fixed datasets. For example, reinforcement learning algorithms interact with their environment, establishing a feedback loop between the learning agent and the training process. A detailed exploration of these algorithms is beyond the scope of this chapter.

While the concepts of unsupervised learning and supervised learning may not have strict boundaries, they provide a general framework to understand the challenges encountered in research and develop new algorithms based on existing models.

2.2 Evaluation and Model Selection

With the rapid advancement of machine learning, several mature tools have arisen to assist researchers in training machine learning models. No longer is training a

machine learning model a tough task. However, it is still essential to analyze the performance of the derived models and choose the most appropriate one. In this section, we will briefly introduce some related concepts and practices.

2.2.1 *Overfitting, Underfitting, and Model Capacity*

One of the key challenges in machine learning is ensuring that our algorithms perform well on new, unseen inputs, rather than just on the training set. The ability to generalize effectively to previously unobserved data is crucial. In this section, we will explore the concepts of overfitting, underfitting, and model capacity, which play a vital role in achieving good generalization.

To begin, we distinguish between the model's error on the training set, referred to as the training error or empirical error, and the error on new samples, known as the generalization error. While our goal is to minimize the generalization error, we can only directly observe the training error since the true nature of new samples is unknown. Consequently, we aim to minimize the empirical error, assuming that it correlates with the generalization error. The field of statistical learning theory provides valuable insights into assessing a model's generalization ability based on assumptions about the data-generating process.

The data used for training and testing is typically generated by a probability distribution known as the data-generating process. Under the assumption of independently and identically distributed (i.i.d.) instances, we consider both the training set and the test set to be drawn from the same distribution. This assumption allows us to model the data-generation process using a probability distribution over individual examples, denoted as p_{data} . By analyzing the relationship between training error and test error within this probabilistic framework, we can gain insights into generalization performance.

While it is theoretically expected that the expected training error of a randomly selected model equals its expected test error, real-world models are not randomly produced. The error observed on a small dataset is merely an estimate of the expected error, resulting in the expected test error often being higher than the training error during model selection. Consequently, when aiming to build a model with good generalization capabilities, two primary considerations come into play:

- Reduce training error.
- Reduce the gap between training and test error.

These two factors correspond to the main challenges in machine learning, namely underfitting and overfitting. Underfitting occurs when the model fails to capture the underlying patterns in the data, resulting in significant training errors. On the other hand, overfitting arises when the model captures noise or idiosyncrasies in the training data, leading to a large gap between the training and test error.

Underfitting and overfitting can have various causes. Overfitting often occurs when the model's capacity, i.e., its ability to fit a wide range of functions, is

excessively high, enabling it to learn intricate patterns present in the training data that may not generalize well. Underfitting, on the other hand, is typically caused by a lack of learning capacity in the model. Modulating the model's capacity can help alleviate the effects of these issues to some extent.

One way to control the model's capacity is by selecting an appropriate hypothesis space, which represents the collection of functions from which the learning algorithm can choose a solution. For instance, linear regression models restrict the hypothesis space to linear functions, while extended linear regression includes polynomial functions, expanding the model's capacity. Statistical learning theory provides several methods for assessing model capacity, with the Vapnik-Chervonenkis dimension [13] being one notable example. However, a detailed discussion of this topic falls outside the scope of this chapter.

In summary, underfitting is relatively straightforward to identify and address since the training error is readily observable, and most machine learning techniques offer ways to increase the model's capacity. Overfitting, on the other hand, presents a greater challenge in real-world applications, and our goal is often to mitigate or reduce its risk rather than completely eliminate it.

2.2.2 Bias and Variance

The bias-variance decomposition is a method for assessing the predicted generalization error of a learning algorithm with regard to a certain task. It attempts to decompose the expected generalization error of the learning algorithm.

Even though all training sets are drawn from the same distribution, it is known that the same algorithm may provide different results on various training sets. As an example, consider the regression task. Let x be test samples, y be ground-truth labels and y_D be collected labels. Let $f(x; D)$ be the model trained on the training set D . The expected prediction of this model is

$$\bar{f}(x) = \mathbb{E}[f(x; D)]. \quad (2.1)$$

Then the expected generalization error of this model can be decomposed as:

$$\mathbb{E}(f; D) = \mathbb{E}_D \left[(f(x; D) - y)^2 \right] \quad (2.2)$$

$$= \text{bias}^2(x) + \text{var}(x) + \epsilon^2 \quad (2.3)$$

where $\text{bias}(x)$ is the difference between expected output and ground-truth

$$\text{bias}(x) = |\bar{f}(x) - y|, \quad (2.4)$$

$\text{var}(x)$ is the variance produced by using different training sets

$$\text{var}(x) = \mathbb{E}_D \left[(f(x; D) - \bar{f}(x))^2 \right], \quad (2.5)$$

and noise ϵ

$$\epsilon = \mathbb{E}_D [|y - y_D|]. \quad (2.6)$$

These three terms characterize the performance of the model from different perspectives.

- *Bias* refers to the deviation between the expected prediction of a learning algorithm and the actual result. It describes the fitting ability of the learning algorithm itself. For example, when a learning method designed for linear models is used to approximate a non-linear function $f(x)$, there will be errors in the estimates due to this simplification.
- *Variance* measures the change in performance caused by variations in datasets of the same size. It captures the impact of data perturbations on the model's predictions.
- *Noise*, also known as irreducible error, represents the lower bound of the generalization error that any learning algorithm can achieve for the given task. It characterizes the inherent difficulty of the problem itself and cannot be reduced by the learning algorithm.

This decomposition demonstrates that the generalization performance is jointly dependent on the algorithm's learning ability, the amount of data, and the complexity of the learning task itself.

The bias-variance trade-off is a crucial issue in machine learning. Ideally, we would like to select a model that accurately represents patterns in the training data and generalizes well to unseen data. However, achieving both simultaneously is often impossible. High-variance learning methods may fit the training set well but are prone to overfitting noisy or unrepresentative training data. On the other hand, algorithms with high bias tend to generate simpler models that may fail to capture important data regularities (i.e., underfitting).

In practical applications, it is desirable to control the degree of training of the learning algorithm for a given learning task. Initially, when the model has not yet been fitted to the training data, even a slight perturbation in the training data can result in significant changes in the model. Therefore, bias dominates the generalization error. As the training progresses, the model gradually fits the training data, and the variance starts to dominate the generalization error. If the training continues excessively, the model may start capturing minor fluctuations or non-global features in the data, leading to overfitting.

2.2.3 *The No Free Lunch Theorem*

The concept of the No Free Lunch Theorem holds significant importance in the field of machine learning, highlighting the fact that no single algorithm can be considered the best for all datasets and situations. [15] investigates whether it is possible to obtain useful theoretical results from training datasets and learning algorithms without making any assumptions about the target variables. It proves that, given a noise-free dataset and a machine learning algorithm, where the cost function is the error rate, the learning algorithm is equivalent when all algorithms are evaluated with the generalization error. This shows that for any two algorithms A and B, even if A outperforms B in many cases, there will be some cases in which B will outperform A. This even holds when one of the algorithms is simply a random guess.

The reason behind this lies in the fact that nearly all machine learning algorithms incorporate certain assumptions regarding the relationship between predictors and target variables. These assumptions introduce what is known as inductive bias into the model. Consequently, different algorithms possess varying degrees of suitability for specific datasets based on the assumptions they make.

The No Free Lunch Theorem serves as a reminder that the selection and development of an appropriate machine learning model necessitate leveraging the available data and domain expertise. There exists no universally superior machine learning algorithm that can effectively tackle all tasks.

2.2.4 *Regularization*

One commonly employed approach for model selection is regularization. Regularization serves as an implementation of the structural risk minimization strategy, often accomplished by incorporating a regularization or penalty term into the empirical risk. This regularization term typically exhibits a monotonic increase in relation to the complexity of the model. In other words, as the model becomes more complex, the corresponding penalty grows.

Generally, the regularization formulation can be expressed as follows:

$$\min_{f \in \mathcal{F}} \quad \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda R(f) \quad (2.7)$$

where the first term is the empirical risk, the second term is the regularization term, and the parameter λ is used to control the ratio of the two. The regularization term can assume various forms, such as the norm of the model parameter vector, including the L_1 norm, L_2 norm, and others. Each form of regularization implies different assumptions about the underlying model. Through the introduction of these

assumptions, the objective is to decrease the model's generalization error rather than solely focusing on minimizing the training error.

Regularization aligns with the principles of Occam's razor, which suggests that among numerous plausible models, the one that effectively explains the available data while maintaining simplicity is typically the preferred choice. In a Bayesian context, the regularization term reflects prior knowledge concerning the model.

2.2.5 Parameter Tuning and Validation

The majority of machine learning algorithms incorporate hyperparameters that govern their behavior. Unlike the model's parameters, these hyperparameters are not learned directly from the training set as they control the model's bias or capacity. Optimizing these hyperparameters directly often leads to overfitting the model to the training set.

To select appropriate hyperparameters, it is common to introduce a validation set that remains unseen by the training algorithm. In a previous section, we discussed the use of a test set, consisting of samples from the same distribution as the training data, to evaluate the model's generalization error after the learning process. However, it is crucial to emphasize that the test samples must not influence model selection, including the selection of hyperparameters. Therefore, the test set should not be used for the validation set. Instead, we typically construct the validation set using samples from the training set.

Specifically, we learn the model parameters from the training set, choose suitable hyperparameters based on their performance on the validation set, and then evaluate the final model on the test set.

It is important to note that the performance on the validation set does not necessarily reflect the model's general performance. While the validation set contains unseen data for the model, the estimation of generalization error based on the performance of the validation set tends to be higher due to information leakage caused by the hyperparameter selection process.

2.3 Supervised Learning Algorithms

Supervised learning is a widely used paradigm in the field of machine learning. It encompasses machine learning systems that are trained on labeled data, where each data point is associated with a known label or outcome. Additionally, unsupervised learning algorithms are often influenced by concepts and techniques from supervised learning. In this section, we will provide a concise introduction to two fundamental supervised learning algorithms that are closely related to subsequent deep learning models. Due to constraints in space, we will only provide a brief overview of these algorithms.

2.3.1 Linear Models

Linear models aim at learning a decision function that makes predictions from linear combinations of features. The model takes the form:

$$f(x) = w^T x + b \quad (2.8)$$

where weight w and bias b are learnable parameters.

Despite their simplicity, linear models can effectively address a wide range of problems. They encapsulate many fundamental concepts in machine learning, and numerous more intricate models can be viewed as extensions or variations of linear models. Furthermore, the interpretability of linear models is noteworthy, as the parameter vector w provides intuitive insights into the contribution of each feature towards the prediction.

For regression tasks, mean squared error (MSE) is the most commonly used performance measure. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^n$, solving a linear regression model can be formulated as the following optimization problem:

$$\underset{w, b}{\operatorname{argmin}} \sum_{i=1}^n (f(x_i) - y_i)^2. \quad (2.9)$$

We can use the least squares method for parameter estimation of the parameters w and b . By making the derivative of the loss function with respect to the parameter to be zero, we can obtain the closed-form solution of the optimal solution for the parameter.

For simplicity, we combine w and b into vector form $\beta = (w; b)$, correspondingly representing the dataset as a matrix X and label y .

This optimization problem can be written as

$$\underset{\beta}{\operatorname{argmin}} (y - \beta^T X)^T (y - \beta^T X). \quad (2.10)$$

Differentiate the loss function with respect to β yields

$$\frac{\partial E_\beta}{\partial \beta} = 2X^T(\beta^T X - y). \quad (2.11)$$

The closed-form solution of the optimal solution of the parameter β can be obtained by setting the above formula to zero. When $X^T X$ is a positive definite matrix, the estimation of parameter β is

$$\beta^* = (X^T X)^{-1} X^T y. \quad (2.12)$$

When $X^T X$ is not a full-rank matrix, there will be multiple solutions. The final choice of model is usually determined by the inductive preference of the learning algorithm, which can be achieved by introducing an appropriate regularization term.

More generally, we consider a monotonically differentiable function $g(\cdot)$, the generalized linear model can be defined as

$$y = g(w^T x + b) \quad (2.13)$$

where $g(\cdot)$ is called the link function.

For a binary classification task, a linear regression model cannot directly produce a 0/1 prediction. An immediate idea is to use a function to convert the predicted real values z to 0/1 values. The most ideal mapping is a step function

$$y = \begin{cases} 0, & z < 0 \\ 0 \text{ or } 1, & z = 0 \\ 1, & z > 0 \end{cases} \quad (2.14)$$

Obviously, the step function is not continuous, which is not conducive to the solution of the linear model. A widely used surrogate is the logistic sigmoid function

$$y = \frac{1}{1 + e^{-z}}. \quad (2.15)$$

Bringing the logistic sigmoid function into the generalized linear model as a link function, we have

$$\ln \frac{y}{1 - y} = w^T x + b. \quad (2.16)$$

Here, y represents the probability that the sample x belongs to the positive class. It is noteworthy that the model employs the prediction output of the linear regression model to approximate the logarithmic probability of the true label. This particular model is commonly referred to as logistic regression.

We can use the maximum likelihood method for parameter estimation. The log-likelihood can be written as

$$l(w, b) = \sum_{i=1}^n \ln p(y_i | x_i; w, b). \quad (2.17)$$

Similarly, let $\beta = (w; b)$ and $x' = (x; 1)$, the optimization problem can be written as

$$\operatorname{argmin}_{\beta} l(\beta) \quad (2.18)$$

where

$$l(\beta) = \sum_{i=1}^n \left(-y_i \beta^T x'_i + \ln(1 + e^{\beta^T x'_i}) \right). \quad (2.19)$$

Since the optimization objective is a derivable convex function, according to the theory of convex optimization, the optimal solution β^* can be obtained by any classical numerical optimization algorithms such as gradient descent.

2.3.2 Support Vector Machine

The Support Vector Machine (SVM) [2] is one of the most influential methods in the field of supervised learning, which is based on the linear model $w^T x + b$. In the context of classification tasks, the SVM aims to identify a hyperplane that effectively separates the training samples while maximizing the “margin” between the hyperplane and the samples. Mathematically, this objective can be formulated as a constrained optimization problem:

$$\max_{w,b} \frac{2}{||w||} \quad (2.20)$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1, \quad i = 1, 2, \dots, n \quad (2.21)$$

SVM encompasses several important concepts, including the kernel trick. It is noteworthy that the linear function within SVM can be expressed as follows:

$$w^T x + b = b + \sum_i \alpha_i x^T x_i, \quad (2.22)$$

where x_i is the training sample, α_i are some coefficients. By rephrasing it in this manner, we can substitute x with the output of a designated feature function $\phi(x)$. Consequently, the dot product is replaced by the kernel function known as $k(x, x_i) = \phi(x) \cdot \phi(x_i)$. The kernel trick’s potency lies in its ability to generate a nonlinear model in the original space by learning a linear model in the newly transformed space, especially when the feature function $\phi(\cdot)$ is nonlinear. The Gaussian kernel, commonly employed as a kernel function, can be defined as follows:

$$k(u, v) = \mathcal{N}(u - v; 0, \sigma^2 I) \quad (2.23)$$

Due to the fact that the Gaussian kernel corresponds to a dot product in an infinite dimensional space, it significantly enhances the capacity of linear models. This

kernel is also referred to as the radial basis function (RBF) kernel, as its value decreases in the outward direction from u towards v .

2.4 Neural Networks and Deep Learning

Machine learning algorithms have demonstrated remarkable performance in various significant tasks, yet they face challenges when dealing with high-dimensional data, which is commonly referred to as the curse of dimensionality. Traditional machine learning methods struggle to generalize and learn complex functions effectively in such spaces, often incurring substantial computational costs. Deep learning (DL) emerges as a solution to address these challenges.

The transition from traditional machine learning to deep learning occurs naturally, as most machine learning models can be seen as a combination of key components: a specific dataset, a cost function, an optimization procedure, and a model. For instance, the linear regression algorithm incorporates a dataset composed of input features X and corresponding targets y , a cost function represented by

$$L(w, b) = -\mathbb{E}_{x, y \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(y | x), \quad (2.24)$$

a model specification

$$p_{\text{model}}(y | x) = \mathcal{N}(y; x^T w + b, 1), \quad (2.25)$$

and an optimization algorithm that often involves solving the normal equations to find the point where the cost gradient is zero.

By recognizing that these components can be replaced independently of each other, we can derive a vast array of algorithms. This is precisely the approach employed in deep learning. We can design different cost functions tailored to specific tasks, typically incorporating terms that facilitate statistical estimation of the learning process. The most common cost function is the negative log-likelihood, which leads to maximum likelihood estimation when minimized. In deep learning, the model is replaced with nonlinear networks, rendering many cost functions intractable to optimize in closed form. This necessitates the adoption of iterative numerical optimization algorithms such as gradient descent.

2.4.1 Feedforward Networks

A prominent deep learning model is the feedforward neural network, also known as the multi-layer perceptron (MLP). In fact, most of the prominent designs in the deep learning community today can be seen as specialized versions or variants of feedforward neural networks.

In essence, a neural network, like other ML models, aims to approximate a function. For instance, suppose that there exists a theoretically optimum classifier f^* that can translate all input samples x to the proper category for a certain classification task. The neural network establishes a model $y = f(x; \theta)$ that approximates the optimum classifier by estimating θ with the training dataset.

Typically, feedforward neural networks can be viewed as compositions of basic functions. If we consider each function as a node in a graph, a model can be defined by this graph, where the topology specifies how these nodes are interconnected to form the model. The term “feedforward” refers to the flow of information within the model. Given an input x , the necessary computations to define f are performed sequentially until the output y is generated. In this process, information flows in a single direction, and the relationship between the output and input is unidirectional. From a topological perspective, the computational graph of the model forms a directed acyclic graph (DAG). The chain structure is the most common architecture in neural networks. For example, if we have three sequentially connected functions, f_1 , f_2 , and f_3 , the network can be expressed as $f(x) = f_3(f_2(f_1(x)))$. In this case, we often refer to f_1 as the first layer, f_2 as the second layer, and so on. The number of functions in this chain determines the depth of the model. Since the typical representation of each intermediate layer in a network is often in the form of vectors or matrices, we refer to the dimensions of these layers as the breadth of the model. The output layer is the final layer of the network. During neural network training, we aim to make the model’s output $f(x)$ match the desired output of the ideal model $f^*(x)$. Naturally, we do not have access to the ideal model in reality, so we collect a large amount of noisy data points as approximations of f^* . Throughout the training process, it is important to note that we do not explicitly specify the behavior of the intermediary layers. The learning algorithm must determine how to utilize these layers to achieve the desired output, as the training data does not specify the function of each layer. Consequently, these layers are referred to as hidden layers since the training data does not provide the intended output for each of them. This training technique is also known as end-to-end learning.

2.4.1.1 Perceptron

The concept of neural networks draws inspiration from observations of the human central nervous system. At the core of a neural network lies the MP neuron, serving as its fundamental component. One can view a neural network as a network of interconnected neurons. The MP neuron model was initially developed to mimic biological neurons and comprises the following elements:

- Inputs x_i simulate the stimulation received by the cell from other nerve cells. Greater input values indicate stronger stimulation.
- Weights w_i represent the sensitivity of the cell to stimuli from different sources.
- The threshold θ describes the activation difficulty of the neuron. A larger threshold requires a higher level of stimulation to activate.

- The activation function $y = f(x)$ determines the neuron's output.

Formally speaking, this neuron model can be written as

$$y = f \left(\sum_{i=1}^N w_i x_i - \theta \right). \quad (2.26)$$

In MP neurons, the activation function takes the form of a step function. When the stimuli surpass the threshold, the neuron outputs 1; otherwise, it outputs 0.

The simplest neural network, called a Perceptron, consists of a single layer of neurons. It defines the following mapping from input to output space:

$$f(x) = \text{sgn}(w \cdot x + b) \quad (2.27)$$

where w and b are learnable parameters and sgn is the sign function.

It is worth noting that the Perceptron is essentially a linear classification model. In fact, generalized linear models can be considered the simplest form of neural networks. The linear regression model and logistic regression model discussed in the previous section can both be regarded as single-layer neural networks, utilizing mean square error and log-loss as their respective loss functions. Similarly, an SVM can be seen as a single-layer network employing the hinge loss function.

2.4.1.2 Multi-layer Perceptron

From a biological perspective, most organisms consist of multiple cells. When a single layer of MP neuron is insufficient, it becomes natural to consider adding more neurons to the Perceptron in order to enhance its power.

Consider a perceptron model with two layers of neurons. If the step function is used as the activation function, the input of the second layer will become discrete values that limit the expression of the model, which is obviously not what we expect. An idea is to use a continuously differentiable function to approximate the step function, a commonly used function is the sigmoid function $S(x) = \frac{1}{1+e^{-x}}$. So far we have obtained a multi-layer perceptron.

At first glance, it may seem that this model offers little improvement over a single-layer perceptron. However, neural networks with hidden layers actually provide a framework for universal approximation. Specifically, the universal approximation theorem states that a feedforward network with a linear output layer and at least one hidden layer, utilizing any “squashing” activation function (such as the sigmoid function), can approximate any Borel measurable function from one finite-dimensional space to another with any desired nonzero amount of error, provided that the network has a sufficient number of hidden units. Moreover, the derivatives of the feedforward network can also approximate the derivatives of the function arbitrarily well. While the concept of Borel measurability is beyond the scope of

this chapter, it suffices to say that any continuous function on a closed and bounded subset of \mathbb{R}^n is Borel measurable and thus can be approximated by a neural network.

The universal approximation theorem validates the immense capacity of feedforward neural networks. A feedforward network with a single hidden layer is capable of representing almost any function. Although the theorem assures us that there exists a network large enough to achieve any desired precision, it does not specify the size of such a network. In fact, in the worst case, the layers of the network may need to be unrealistically large, hindering effective learning and generalization. In many cases, employing a network with multiple layers is preferred as it reduces the overall number of model parameters, making the learning and generalization process easier.

2.4.2 *Back Propagation of Errors*

In the realm of neural networks, the back-propagation algorithm [11] plays a pivotal role in enabling effective training and optimization. As a neural network processes input data and produces output predictions, information flows forward through the network in what is known as forward propagation. However, the ultimate goal of training is to minimize the discrepancy between the predicted output and the desired output, typically represented by a scalar loss function.

The back-propagation algorithm acts as a conduit for the flow of information in the opposite direction, allowing the cost function's information to propagate backward through the network. By doing so, it facilitates the calculation of parameter gradients, which are crucial for optimizing the network's performance. Through the utilization of these gradients, the network's parameters can be adjusted iteratively to minimize the loss function, ultimately leading to improved predictions.

Conceptually, a deep feedforward neural network can be thought of as a composite model, consisting of multiple interconnected layers, each comprising linear transformations and activation functions. By applying the chain rule from calculus, we can compute the partial derivative of the loss function with respect to each model parameter, layer by layer, starting from the cost function. This process is commonly referred to as the back-propagation of errors.

The back-propagation algorithm takes advantage of the differentiability of the individual components within the network. Given that all operations in a neural network are differentiable, the composite function formed by these operations is also differentiable. By invoking the chain rule on the computational graph, the back-propagation algorithm efficiently calculates the gradient of each parameter in the network, layer by layer, in a recursive manner.

At each layer of the network, the algorithm calculates the local gradient, which represents the sensitivity of the layer's output to changes in its input. This local gradient is then combined with the gradients from subsequent layers, weighted by the respective parameters connecting them. This process is akin to “back-

propagating” the error signals from the output layer to the input layer, hence the name of the algorithm.

The calculated gradients are subsequently used to update the network’s parameters using an optimization algorithm, such as stochastic gradient descent (SGD) or one of its variants. By iteratively adjusting the parameters based on the gradients, the network gradually converges to a state where the loss function is minimized, leading to improved performance and more accurate predictions.

While the specific derivation of the back-propagation algorithm is beyond the scope of this chapter, it is worth noting that numerous resources are available that delve into the mathematical details. Interested readers are encouraged to explore these references to gain a comprehensive understanding of the algorithm’s inner workings and its mathematical foundations.

2.4.3 *Architecture Design of Neural Networks*

The development of deep learning has emphasized the significance of designing neural network architectures tailored to specific tasks. In this context, a neural network can be seen as a composition of simple functions, referred to as “layers”. Regardless of the complexity of the neural network architecture, as long as the layers are differentiable, the standard back-propagation algorithm and gradient-based optimization techniques can be employed for model training. Many applications employ highly intricate models consisting of multiple modules. Therefore, this section aims to introduce commonly used building blocks rather than focusing on a specific neural network architecture.

- **Linear layers:** Linear layers, also known as fully connected layers, serve as fundamental units in neural networks. Each node in a linear layer is connected to all nodes in the preceding layer and is responsible for synthesizing previously extracted features. Due to their fully connected nature, these layers encompass a large number of parameters. Linear layers excel at capturing complex relationships between input features and can learn intricate patterns in the data. Typically, the final layer of most models is a fully connected layer that maps the learned representation to the shape of the target.
- **Convolution layers:** Convolution layers are essential components of Convolutional Neural Networks (CNNs) [5, 12]. They can be viewed as specialized versions of fully connected layers. By utilizing filters of a specific size, convolution layers convolve the input data and produce multiple feature maps as output. This filtering process allows the network to extract local patterns and spatial information from the input. Parameter-sharing within the convolution layer reduces the number of parameters compared to fully connected layers, which helps mitigate the risk of overfitting and accelerates network computation. Convolution layers are particularly effective in processing grid-like data such as images and are widely employed in computer vision applications.

- **Recurrent Layer:** Recurrent layers are vital elements of Recurrent Neural Networks (RNNs) [11]. Unlike other layers, recurrent layers have connections between nodes that create cycles, allowing the output from certain nodes to influence subsequent input to the same nodes. This cyclic nature enables recurrent layers to exhibit temporal dynamic behavior, making them highly suitable for processing sequential and time-series data. RNNs, including variants such as the Long short-term memory (LSTM) network [6], excel at capturing dependencies over varying time scales. LSTMs are specifically designed to mitigate the vanishing gradient problem in traditional RNNs, allowing them to retain important information for long periods and effectively model sequences.
- **Transformer:** The Transformer model [14], which has emerged as a powerful architecture in recent years, represents a paradigm shift in sequence modeling, particularly in the field of Natural Language Processing (NLP). At the heart of the Transformer is the self-attention mechanism, which allows the model to selectively weigh the importance of each part of the input sequence for different samples. Unlike recurrent layers, the Transformer processes the entire input sequence simultaneously, enabling parallel computation and capturing long-range dependencies effectively. By attending to different parts of the input at different stages, the self-attention layer enables the model to capture complex patterns and relationships across the sequence. The Transformer architecture has demonstrated remarkable success in various NLP tasks, including machine translation, text generation, and sentiment analysis.
- **Normalization layers:** Normalization layers encompass a group of methods utilized to expedite and stabilize the training of artificial neural networks [7]. These methods achieve this through the normalization of inputs to the layers, achieved by re-centering and re-scaling. Batch Normalization and Layer Normalization are widely employed options in this regard. BatchNorm normalizes the input across a mini-batch, reducing the internal covariate shift and improving the generalization of the model. LayerNorm, on the other hand, normalizes the input within each layer, making the model less sensitive to the scale of inputs and improving its robustness.
- **Non-linear activations:** Activation functions serve as fundamental components of neural networks, introducing nonlinearity into the model. Nonlinearity is crucial for the network to capture complex relationships and make the model capable of learning intricate patterns. Rectified Linear Unit (ReLU) is presently the most commonly used activation function due to its simplicity and non-saturating properties. ReLU sets all negative input values to zero and preserves positive values, allowing the network to easily learn sparse representations. Bounded activation functions such as sigmoid and hyperbolic tangent (tanh) are also widely used, particularly in classification tasks, as they can map arbitrary data to a specific range and provide probabilistic interpretations.

2.5 Concluding Remarks

Machine learning and deep learning have witnessed exponential growth and have become integral to various multidisciplinary research fields, including the computational modeling of aesthetics, emotion, and artistic style. This chapter has served as a stepping stone for researchers who possess limited knowledge or are new to the ML community, providing them with essential background information.

Throughout this chapter, we have defined learning algorithms and explored fundamental concepts and principles underlying machine learning systems. Additionally, we have offered a concise introduction to several typical and basic machine learning models. Our deliberate omission of extensive formula derivations was intended to cater to novice readers, particularly those without formal training in computing, machine learning, or statistical modeling.

It is crucial to acknowledge that the scope of this chapter is limited, and we could not cover all aspects of this rapidly evolving field. However, we hope that by offering an overview, we have ignited your curiosity and provided a solid foundation to facilitate further exploration. To delve into cutting-edge advancements and more comprehensive coverage of the fundamentals, we encourage you to refer to relevant textbooks [1, 3, 4, 9, 10] and engage with recent research papers.

Machine learning and deep learning continue to push the boundaries of knowledge and drive innovation across numerous disciplines. As you embark on your journey, we urge you to embrace the dynamic nature of this field and be inspired by its potential to revolutionize research and industry alike. Stay curious, stay engaged, and keep pushing the boundaries of what is possible through the power of machine learning and deep learning.

Acknowledgments The work was funded in part by a generous gift from Amazon to the author's dissertation advisor Professor James Z. Wang. The author also acknowledges the advice and constructive comments from him.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin/Heidelberg (2006)
2. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995)
3. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016). <http://www.deeplearningbook.org>
4. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, Berlin (2009). <https://books.google.com/books?id=eBSgoAEACAAJ>
5. 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, pp. 770–778 (2016)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>

7. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning (PMLR), pp. 448–456 (2015)
8. Mitchell, T.M.: Machine Learning, 1st edn. McGraw-Hill, Inc., New York (1997)
9. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. The MIT Press, Cambridge (2012)
10. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. The MIT Press, Cambridge (2012)
11. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *Nature* **323**(6088), 533–536 (1986)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Preprint (2014). arXiv:1409.1556
13. Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16**(2), 264–280 (1971). <https://doi.org/10.1137/1116025>
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017)
15. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. *Neural Comput.* **8**(7), 1341–1390 (1996). <https://doi.org/10.1162/neco.1996.8.7.1341>

Part II

Human Social Vision

This part is focused on how the human visual system reads social and emotional information from the human face, body, and pictorial scenes, and highlights ethical issues associated with extending this work into the domain of computer vision.

The first chapter, “Facing a Perceptual Crossroads: Mixed Messages and Shared Meanings in Social Visual Perception,” explores the combinatorial nature of social visual perception and offers a functional account for how various cues from the face (e.g., emotion, gender, race, eye gaze) combine and interact in human perception to form the type of holistic social meaning from which we derive impressions of others.

The second chapter, “Social Vision of the Body In Motion: Interactions Between the Perceiver and the Perceived,” similarly examines compound social cue processing of human bodies, emphasizing influences of both the perceiver and the target of perception.

The third chapter, “Visual Perception of Threat: Structure, Dynamics, and Individual Differences,” extends this further to examining threat perception from faces, bodies, and the broader context by examining how the human visual system perceives threat from pictorial scenes. The neural mechanics and dynamics underlying the processing of ambiguous versus clear threat-related cues are detailed in an adaptive dual visual pathway model.

Finally, the fourth chapter, “From Pixels to Power: Critical Feminist Questions for the Ethics of Computer Vision,” delineates important ethical issues drawn from psychological and feminist theory to consider when extending human social visual perception into the realm of computer vision.

Chapter 3

Facing a Perceptual Crossroads: Mixed Messages and Shared Meanings in Social Visual Perception



Natalie Strand, Nicole Hedgecoth, and Reginald B. Adams, Jr.

Abstract The human body conveys social information through a myriad of cues, one of the most important being emotion. Perception of emotion in faces has been extensively researched, with a particular focus on the impact of eye gaze on emotion recognition. Within face perception work, the social functional approach was developed to explain how combined processing of social cues (e.g., emotion, eye gaze, race, gender/sex) is highly adaptive and necessary to facilitate social interaction. Recently, computer models and machine learning have been incorporated into person perception research to gain a clearer understanding of how physiognomic cues interact with emotion information to inform emotion perception. In this chapter, we review past work examining emotion perception at the intersection of race and gender/sex and highlight important findings from human perceivers and computer models. We also address important ethical considerations involved in using new technology to conduct social perception research and emphasize the critical role of Intersectionality.

3.1 Introduction

How we process and perceive each other visually through the face and body, eye gaze, emotion expression, race, gender/sex, appearance, and age was long considered due to independent processes, and thus each type of social cue was often isolated within separate fields (see [8]). Critically, the combined and interactive influences of these cues on social perception were, therefore, not adequately addressed by prior research or theory.

For example, in the vision science domain, face processing models (see [28, 53]) long argued that social cues conveyed by the face, such as identity, appearance, eye gaze, and emotion expression, are extracted in a noninteracting and independent

N. Strand (✉) · N. Hedgecoth · R. B. Adams, Jr.

Department of Psychology, The Pennsylvania State University, University Park, PA, USA

e-mail: njs6003@psu.edu; nph5199@psu.edu; rba10@psu.edu

manner (see also [72]). These assumptions, however, do not align with current notions about how the human brain evolved to extract predictions about what others around us are thinking and feeling (e.g., [4, 12, 15] see also [36, 37]). Arguably, it is only in combination that social cues convey adaptive information about the internal states of others, such as their wishes, desires, feelings, and intentions. Given the combinatorial nature of person perception, it makes better adaptive sense that our visual system would have evolved to integrate multiple cues when extracting higher-level social impressions of others. To understand social perception in this way, we must first consider what information we seek to perceive in others in the first place. This framework puts social meaning derived from nonverbal cues as an essential function of the social visual system [8]. Thus, a dominant looking (low brow, thin lips, angular) face staring at you, expressing a frown should enhance threat detection compared to seeing that same anger expression on a submissive looking (high brow, full lips, round) face looking down. Unlike previous source-input models of face perception (e.g., focused on independent processing of expression, gaze, identity, appearance), this social functional approach (see [4, 8, 12, 94]) focuses on the underlying meaning conveyed by these cues and their combined ecological relevance to the observer. In this way, combined processing of social cues is not considered maladaptive as prior theories suggested, but highly adaptive. Only in combination do social cues perceptually inform the unified representations that guide our impressions of and responses to others.

3.2 The Role of Eye Gaze in Emotion Perception

A link between eye gaze behavior and emotional expression is readily apparent in our language. We “stare” another down when angry, avert eyes in disgust, dart them away in fear, display downcast eyes when sad, gaze longingly into a lover’s eyes. Given that we regularly associate different eye gaze behavior with the experience and expression of various emotions, it stands to reason that the perception of emotion would also be influenced by a concurrent display of such eye gaze behavior. Yet, for many years this possibility remained an untested proposition (see [2, 6]).

The reason for this is likely due to early, prominent eye gaze researchers explicitly assuming eye gaze direction was unlikely to affect an observer’s perception of a concurrently displayed facial affect [41]. This contention fits with contemporary face processing models that suggest functionally distinct sources of visual information are processed via functionally distinct neural processing routes, including different visual pathways underlying gaze and expression perception (e.g., [28, 29, 52]). Others shared the generally accepted but untested supposition that direct eye gaze should impact how we interpret all emotional output, as eye contact communicates to an observer that he or she is the object of another’s attention and therefore the target of whatever emotion is being displayed, making facial expression self-relevant to the observer see also [5]. This assumption predicts

that direct relative to averted eye gaze behavior will enhance the communicated intensity of emotion regardless of type of expression displayed [18, 40, 68, 93].

Some work has supported this sort of main effect of eye gaze in emotion perception (i.e., as predicted by the direct-gaze attention capture hypothesis). Direct eye gaze holds attention to a greater degree than averted displays [45], whereas averted gaze tends to shift attention away from the face [35, 71]. Based on such effects of visually mediated attention, a logical prediction is that direct eye gaze would enhance the processing of all emotional displays regardless of any shared signals. Indeed evidence has been found to support this contention such that direct relative to averted eye gaze has been found to facilitate the processing of emotion recognition (see [21, 48, 67]). This finding fits with a number of other studies showing a generalized influence of direct eye gaze on the processing of all forms of face perception (see [88]); but see also [92] for evidence of enhanced gender/sex discrimination for averted versus direct gaze faces. This account suggests an indirect, unitary influence of visually mediated attention on the processing of emotional expression, but one that does not address the now numerous demonstrations of interactivity.

In the last 20 years, significant interdisciplinary attention has been devoted to the issue of combined processing of eye gaze and emotion. This work demonstrates both the complexities and new insights that can result when examining the combinatorial nature of face processing. Although both of these cues are expressive, they nevertheless have been argued to be processed via two distinct neural systems [28]. Mounting evidence, however, supports the interdependency of processing of eye gaze and emotion, using psychophysical, self-report ratings, and neuroimaging paradigms (e.g., [46, 48, 49, 58, 59, 75, 87, 90]), and clear patterns of interactivity have been established [2, 3, 6, 44, 48, 69, 86, 87].

Notably, interdependencies between eye gaze and emotion cannot be attributable to any visually confounded properties, as these two cues occupy nonoverlapping space within the face (i.e., eyes can change direction without influencing facial muscle patterning and vice versa). Below we review recent work conducted on this issue.

In our initial examination of eye gaze and emotion perception, we introduced the “Shared Signal Hypothesis” [2, 3, 7]. This predicts that when paired, cues relevant to threat that share a congruent underlying signal value should facilitate the processing efficiency of threat. In support of this hypothesis, using speeded reaction time tasks and self-reported perception of emotional intensity, Adams and Kleck [2, 3] found that direct gaze facilitated processing efficiency, accuracy, and increased the perceived intensity of facially communicated approach-oriented emotions (e.g., anger and joy), whereas averted gaze facilitated processing efficiency, accuracy, and perceived intensity of facially communicated avoidance-oriented emotions (e.g., fear and sadness). Similar effects were replicated by Sander et al. [86] using dynamic threat displays, and by Hess et al. [56] who found that direct relative to averted anger expressions and averted relative to direct fear expressions elicited more negative responsivity in observers. Further, Mathews et al. [75] found a faster cueing effect for fear faces than neutral faces in participants with high anxiety but

not low anxiety (see also [44, 80]). Additionally, attention capture of direct gaze was greater when displayed with angry expressions than fear or neutral expressions [44]. Finally, when eye gaze was shifted after emotion was presented, fearful faces were found to induce higher levels of cueing compared to other emotions for all participants regardless of anxiety level [90]. These findings again support the idea that relative differences in timing of eye gaze and emotion processing can influence perceptual integration.

Since these first studies, evidence for social functional interactions across an ever-widening array of compound social cues has burgeoned, including (to name a few): (1) gaze and head postures [30], (2) race and emotion [1, 13, 60], (3) race and eye gaze [8, 83, 91], (4) gender/sex and emotion [10, 19, 54, 55]; cf [72], (5) facial expression and approach/avoidance movement [7], and (6) gaze and facial attraction [66]. Further, work has demonstrated that body language [16], visual scenes, and vocal cues also influence the processing of facial displays of threat [76]. All of these cues have been found to influence perception even at the earliest stages of face processing, across conscious and nonconscious processing routes [4, 77]. These findings suggest that the visual integration of compound social cues occurs very early in visual processing, even when the combined cues display divergent perceptual properties (e.g., eye gaze, faces, bodies, voices), yet in combination convey adaptive social affordances.

Dating back to Allport (e.g., [17]), extensive research has examined how we use social categorization (e.g., gender/sex, race, age) to help make sense of others around us see also [26, 27, 42]. In terms of examining how social categories influence emotion perception, most of the work has examined how gender/sex influences emotion perception with the face as a primary vehicle for examining such effects see [10] for review. The influence of race, particularly comparing Black versus White faces, on emotion perception has also received some empirical attention over the last couple of decades, though far less than gender/sex (e.g., [15]). We review each in turn below.

3.3 Gender/Sex and Emotion

A combination of deeply ingrained stereotypes and facial appearance cues influence how people distinguish emotions and form impressions based on gender/sex and emotion-related information. In this section, we review the intersecting nature of gender/sex and emotion and discuss the informative role of computer vision in understanding how gender/sex and emotion impact face and body perception. We first discuss the pervasive gender/sex-emotion expectations that inform emotion perception and then detail how the overlapping nature of gender/sex and emotion related cues in the face inform our understanding of how humans and computers derive social meaning from a face.

Across a range of research examining gender/sex and emotion, there is a shared agreement on a set of stereotypes that are ascribed to the emotional experiences

and expressions of men and women. Not only are women perceived as more emotional than men, but people assume that women experience more emotions at a greater frequency (e.g., [89]). Women are typically expected to feel emotions more intensely than men [84]. While this evidence suggests that people perceive women to generally experience all emotions more than men, there are clear differences in the way men and women are expected to display certain emotions. Women are stereotyped as experiencing more powerless emotions (e.g., happiness, shame), whereas men are expected to exhibit more powerful emotions (e.g., anger, contempt; [79]). This distinction in expected emotions for men and women is also broadly related to stereotypes of women's communal and affiliative nature compared to men's agentic and powerful disposition (see [10]). Furthermore, in line with gender/sex differences in expected emotions, the Stereotype Content Model (SCM) (see [43]) proposes that men are perceived as cold yet competent, while women are perceived as warmer and less competent. However, follow-up research suggests that women and men are relatively equal in terms of perceived competence, with women still maintaining higher levels of perceived communal and men maintaining higher agency [39].

Within the field of face perception research, the intersection of gender/sex and emotion cues has received significant attention. In alignment with gender/sex-emotion expectations, past work has found that facial cues related to femininity and masculinity impact how people respond to certain emotional expressions. When presented with angry and happy faces, people tend to respond faster to angry men versus angry women but also respond faster to happy women versus happy men [19]. Furthermore, people display an attentional bias toward women expressing fear and men expressing anger [54, 55].

There are several arguments aimed at elucidating the differential processing of emotions in the faces of women and men. In particular, Adams et al. [10] proposed that two primary explanations drive our understanding of emotion perception at the intersection of gender/sex and emotion; namely, a stereotype and structural account (see also [31]). First, a set of well-established gender/sex-emotion stereotypes influence how emotions are perceived in women versus men. Second, the structural account details how an overlap of facial cues related to gender/sex and emotion simultaneously influence emotion perception [19, 57]. For instance, certain features consistent with expressions of fear, such as higher eyebrows, are typically associated with facial cues related to femininity, whereas other features, such as low brows, are typically related to both masculinity and anger [9]. This overlap is expanded upon in the Common Cue Hypothesis [13] which describes how emotion-related physiognomic features that are linked to perceptions of social categories (i.e., gender/sex, age, race) influence how we perceive emotion at the intersection of these categories. Given that facial cues related to emotions and femininity/masculinity coincide with gender/sex emotion stereotypes, it is difficult to discern exactly how people perceive emotions displayed by women and men, which extends to research exploring perception of emotion in older adult women and men (e.g., [11]). The literature suggests that structural and stereotype cues are deeply intertwined and likely driving our perceptions of emotion expressions in tandem.

In addition to faces, bodies also signal important emotional information [74]. Indeed, there is evidence to suggest that emotion perception from body posture is often as accurate as emotion recognition from the face [33]. Furthermore, gender/sex stereotypes have also been shown to influence how people decode emotion from the body [20, 64]. For example, past research has found that people are more likely to categorize angry body expressions as men, whereas sad body expressions are categorized as women [64].

Recently, machine learning has been adopted to examine the intersection of gender/sex-related facial cues and emotion expression [14, 97]. In such studies, researchers train image classifiers to distinguish between different emotional expressions and then apply those classifiers to neutral faces. Zebrowitz and colleagues [97] were the first to develop this method. They trained connectionist models to identify different facial expressions of emotion (e.g., happy, surprised); then, after training, the models were presented with neutral faces of women and men to test patterns of overlap in the facial metrics between neutral and emotion faces. The network revealed that structural cues in the neutral faces (e.g., the height of brow) coincided with and mirrored certain emotional expressions. Specifically, their results demonstrate that neutral expressions on men's faces appeared angrier while neutral expressions on women's faces appeared more surprised. Moreover, recent work by Albohn and Adams [14] replicated and extended this research with their own computer vision model. Their model perceived neutral faces of men as resembling anger while those neutral faces of women resembling fear. These recent investigations into the perception of emotion across gender/sex through computer models clearly align with past research exploring the Common Cue Hypothesis with human perceivers. Given that computer models are not subject to the influence of gender/sex stereotypes and provide a more objective explanation of emotion perception, past research employing computer models confirms that human perceivers certainly rely on common cues in the face to form impressions. However, the majority of work exploring emotion perception through computer models has focused specifically on gender/sex cues: future work will need to consider how computer models interpret cues such as age, weight, and their intersection with gender/sex.

3.4 Race and Emotion

Unlike gender/sex, race plays a more complicated role in shaping emotion perception. Within this section, we will cover the literature that centers on race-emotion stereotypes and how these stereotypes intersect with facial-phenotypic expression. We then highlight the apparent contradiction between emotion resembling race-related appearance and expressive cues associated with race-related emotion stereotypes, which contrasts with the research presented in the previous section showing a confound in gender/sex appearance cues and stereotypical emotional expressions.

As with gender/sex, race as a visible social identity evokes different perceptions and emotion stereotypes. Unlike gender/sex, however, less work has been done on examining the perceptions of agency and communalism, though some work does suggest that Black people are perceived as more communal while White people tend to be labeled as more agentic [85]. The SCM suggests that stereotypes of warmth and competence are differentially applied based on the subtypes within the category of Blackness. Other culturally pervasive stereotypes influence emotion perception as well. For instance, Blackness is often associated with being dangerous and criminal [34, 47] and even hypermasculinized [65] which influences emotion stereotypes of being angry (i.e., the ‘angry’ Black person; [38]). Thus, stereotypic associations of anger with Blackness shape this negative characterization of Afrocentric features (e.g., darker skin tone, fuller lips, wide nose; [24]) and can have serious downstream consequences. For example, Blair and colleagues [25] showed that those with more Afrocentric features generally received harsher prison sentences than those with less Afrocentric features.

Some of the initial work in the domain of face perception at the intersection of race and emotion highlight the involvement of these stereotypic effects. Hugenberg and Bodenhausen [62] found, for example, that higher implicit prejudice for White participants against Black people is predictive of noticing the onset and delayed offset of anger on Black faces compared to White faces for White participants. In another set of studies, racially ambiguous faces with angry expressions, compared to happy expressions, were perceived as Black more frequently [62]. Hutching and Haddox [63] replicated those prior findings, as well as extended them to support that the same angry expressions were rated as appearing more intense simply when labeled as Black versus White faces. Critically, these studies used computer generated faces, which have been critiqued for their lack of applicability (see [32]).

More recent work has used both machine learning models and real faces to examine emotion overgeneralization, we find contrasting patterns emerge. Zebrowitz et al. [97] connection models showed that Black faces resembled surprise expressions while White faces resembled anger expressions. Later, Adams et al. [13] used a computer vision model and human raters to replicate similar effects. The machine learning model replicated the effect of White faces structurally resembling anger and extended the structural resemblance of fear expressions to Black faces. Further, human participants also rated fear-resembling neutral faces as more prototypically Black and anger-resembling neutral faces as more prototypically White. Participants also responded faster to fearful Black faces and angry White faces, contrary to previous findings (e.g., [61]). These findings provide more support for the Common Cue Hypothesis on a phenotypic level such that the overlapping cues in emotion and race improve accuracy.

Taken together, the pervasive nature of Black-anger/threatening stereotypes examined in prior research appears to override appearance cues typically associated with race. These stereotypes have serious implications for influencing what we see. For example, Halberstadt and colleagues found that preservice school teachers were generally less accurate at identifying emotions of adult Black faces than White ones and demonstrated an anger bias towards Black faces [50] and replicated those results

with children’s faces [51]. Stereotypes about emotion and race can also be encoded into our machine learning programs and operations of artificial intelligence (AI) systems. For example, AIs that specialize in facial recognition, such as Face++, have been shown to interpret Black faces as expressing more anger, or in the case of Microsoft’s Face API, interpreting more contempt on Black faces, regardless of their equivalently rated smiles [82]. Amazon’s Rekognition exhibited a similarly biased pattern by rating Black men as angrier than White men [70].

As we conclude this section, we acknowledge that the literature we covered almost exclusively pertains to examining racial differences from the dichotomous Black-White approach. We recognize that perceptions of Blackness in tandem with White are not universally applicable to all people of color and should not be taken as generalizable in that regard. For instance, the stereotypes applied to Asian people vary dramatically from those of Black people [47]; even within the identity of “Asian” there are a wide variety of stereotypes that apply heterogeneously (e.g., East Asian and South Asian people are stereotyped differently). Race prototypicality marked through skin-tone also shapes our stereotypes [73]. Each of these differences in stereotypes may influence the emotion overgeneralization on faces from various racial and ethnic backgrounds. As such, we emphasize that future researchers continue to explore the role of stereotypes and perceptions of other racial and ethnic groups to clarify the interplay between stereotype and phenotype on perceptions and, in turn, how that shapes computer vision, machine learning, and AI.

3.5 Other Intersections

3.5.1 Who Else Is Missing?

Along a similar vein, much research has not explored the intersection of race, gender/sex, and emotion. The literature reviewed in the gender/sex and emotion section of this chapter centered on the differences between White women and men; similarly, the following section mainly centered on differences between Black and White men. Given the convergences of stereotypes and perception for gender/sex and emotion but not race and emotion, it is arguably even more important to extend our understanding of perception to these intersections. Further, there is evidence that race is often gendered, with Blackness being masculinized and Asianness being feminized [65, 95]. Additionally, most of the research reviewed here also centers on the perception of younger faces. Young and middle-aged adults are most widely represented and explored in the realm of face perception, to the detriment of older adult faces (see Hedgecoth et al., *in press* for review). This underrepresentation is particularly problematic when we recall the influence of cues like maturity and babyfacedness in face perception.

3.5.2 *Other Identities to Consider? Future Directions*

Beyond the domains that are integral to the work of person perception researchers (e.g., race, gender/sex, and age), new research areas are gaining traction. We will expand on three areas that computer vision researchers should also be especially attuned to: the body, weight, and socio-economic status (SES). We know little about how body and face perceptions compare or how multiple identities influence such perceptions. As such, we encourage researchers to explore the body, especially in tandem with the face, as a cue for understanding emotion perception. For example, Albohn et al. [16], across two studies, showed strong integration effects of shared emotion cues on faces and bodies; bodies were particularly useful for disentangling ambiguity from facial expressions alone. Race and gender/sex both influence perception of bodies, as well. Wilson et al. [96] showed that people perceive Black men to be larger and more physically threatening compared to White men. Other work has also demonstrated that bodily-emotional displays are gender stereotyped (e.g., anger cued for masculinity; [64]). Therefore, bodies are a future direction that researchers should carefully consider, especially in the context of race and gender, as they may offer a more nuanced view of emotion perception.

Implicitly connected to body perception, and should be scrutinized more carefully in face perception work, is the influence of perceived weight. Re and Rule [81] showed that fatness in the face influences impressions of health and attractiveness. Given that fatness is a visible social stigmatizing identity and is also tied with perceptions of both race and gender/sex, future researchers should consider how these intersections affect emotion perception. Similarly, SES has garnered more attention recently in the field of emotion perception. Some evidence suggests that people associate negative emotions (e.g., anger, sadness) with lower SES, while positive emotions (e.g., happiness) were associated with higher SES people [22, 23].

3.6 Conclusions

There are myriad directions for research to flourish and grow in the coming years, with no shortage of questions to be investigated about how different social cues interact with one another to influence our perceptions and impressions. As computer vision, machine learning, and AI programs become more readily available tools to examine such perceptions, there are questions about the ethics and purpose of such tools.

In the realm of face perception research, we must especially be mindful not to reproduce harmful legacies of physiognomy by allowing machine learning to be framed as neutral or objective tools. Particularly since bias is encoded into tools without explicit knowledge. Rather, machine learning and computer vision tools can be used to reveal and convey the biased nature of human social perception. For example, Peterson and colleagues [78], using machine learning and deep neural

networks, created a model that can produce synthetic photorealistic faces that evoke specific attributes in the face. The intention behind the model is to understand the sources of biases (i.e., stereotypes) and how to combat them. But intention alone is not enough. We must be clear in our ethical positions and ensuring that these tools are used to reduce harm rather than reproduce it.

The future of face perception research is exciting, rife with nuanced questions and new tools (e.g., machine learning, AI) to help us further understand human social perception. As the field moves forward, researchers should pay attention to how multiple identities intersect to influence our perceptions of others. And to do so, there must be consideration of who has been historically ignored or missing in research. Further, with the creation of and increased accessibility tools to be reflective of human social perception, they must also be concerned with the ethics of their work and how these tools can be used to reduce harm, especially to historically marginalized people.

References

1. Ackerman, J.M., Shapiro, J.R., Neuberg, S.L., Kenrick, D.T., Becker, D.V., Griskevicius, V., Maner, J.K., Schaller, M.: They all look the same to me (unless they're angry) from out-group homogeneity to out-group heterogeneity. *Psychol. Sci.* **17**(10), 836–840 (2006)
2. Adams, R.B., Jr., Kleck, R.E.: Perceived gaze direction and the processing of facial displays of emotion. *Psychol. Sci.* **14**(6), 644–647 (2003)
3. Adams, R.B., Jr., Kleck, R.E.: Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion* **5**(1), 3 (2005)
4. Adams, R.B., Jr., Nelson, A.J.: Intersecting identities and expressions: the compound nature of social perception. In: *The Oxford Handbook of Social Neuroscience*, p. 394. Oxford University Press, Oxford (2011)
5. Adams, R.B., Jr., Nelson, A.J.: Eye behavior and gaze. In: *APA Handbook of Nonverbal Communication*. American Psychological Association, Washington (2016)
6. Adams, R.B., Jr., Gordon, H.L., Baird, A.A., Ambady, N., Kleck, R.E.: Effects of gaze on amygdala sensitivity to anger and fear faces. *Science* **300**(5625), 1536–1536 (2003)
7. Adams, R.B., Jr., Ambady, N., Macrae, C.N., Kleck, R.E.: Emotional expressions forecast approach-avoidance behavior. *Motivation and Emotion* **30**, 177–186 (2006)
8. Adams, R.B., Jr., Ambady, N., Nakayama, K., Shimojo, S.: The science of social vision: The Science of Social Vision, vol. 7. Oxford University Press, Oxford (2010)
9. Adams, R.B., Jr., Nelson, A.J., Soto, J.A., Hess, U., Kleck, R.E.: Emotion in the neutral face: a mechanism for impression formation? *Cogn. Emot.* **26**(3), 431–441 (2012)
10. Adams, R.B., Jr., Hess, U., Kleck, R.E.: The intersection of gender-related facial appearance and facial displays of emotion. *Emot. Rev.* **7**(1), 5–13 (2015)
11. Adams, R.B., Jr., Garrido, C.O., Albohn, D.N., Hess, U., Kleck, R.E.: What facial appearance reveals over time: when perceived expressions in neutral faces reveal stable emotion dispositions. *Front. Psychol.* **7**, 986 (2016)
12. Adams, R.B., Jr., Albohn, D.N., Kveraga, K.: Social vision: applying a social-functional approach to face and expression perception. *Curr. Dir. Psychol. Sci.* **26**(3), 243–248 (2017)
13. Adams, R.B., Jr., Albohn, D.N., Hedgecoth, N., Garrido, C.O., Adams, K.D.: Angry white faces: a contradiction of racial stereotypes and emotion-resembling appearance. *Affect. Sci.* **3**(1), 46–61 (2022)

14. Albohn, D.N., Adams, R.B., Jr.: The expressive triad: Structure, color, and texture similarity of emotion expressions predict impressions of neutral faces. *Front. Psychol.* **12**, 612923 (2021)
15. Albohn, D.N., Adams, R.B., Jr.: The social face hypothesis. *Affect. Sci.* **3**(3), 539–545 (2022)
16. Albohn, D.N., Brandenburg, J.C., Kveraga, K., Adams, R.B., Jr.: The shared signal hypothesis: facial and bodily expressions of emotion mutually inform one another. *Atten. Percept. Psychophys.* **84**(7), 2271–2280 (2022)
17. Allport, G.W., Clark, K., Pettigrew, T.: *The Nature of Prejudice*. Addison-Wesley, Reading (1954)
18. Argyle, M., Cook, M.: *Gaze and Mutual Gaze*. Cambridge University Press, Cambridge (1976)
19. Becker, D.V., Kenrick, D.T., Neuberg, S.L., Blackwell, K., Smith, D.M.: The confounded nature of angry men and happy women. *J. Pers. Soc. Psychol.* **92**(2), 179 (2007)
20. Bijlstra, G., Holland, R.W., Dotsch, R., Wigboldus, D.H.: Stereotypes and prejudice affect the recognition of emotional body postures. *Emotion* **19**(2), 189 (2019)
21. Bindemann, M., Mike Burton, A., Langton, S.R.: How do eye gaze and facial expression interact? *Visual Cogn.* **16**(6), 708–733 (2008)
22. Björnsdóttir, R.T., Rule, N.O.: The visibility of social class from facial cues. *J. Pers. Soc. Psychol.* **113**(4), 530 (2017)
23. Björnsdóttir, R.T., Rule, N.O.: Negative emotion and perceived social class. *Emotion* **20**(6), 1031 (2020)
24. Blair, I.V., Judd, C.M., Sadler, M.S., Jenkins, C.: The role of afrocentric features in person perception: judging by features and categories. *J. Pers. Soc. Psychol.* **83**(1), 5 (2002)
25. Blair, I.V., Judd, C.M., Chapleau, K.M.: The influence of afrocentric facial features in criminal sentencing. *Psychol. Sci.* **15**(10), 674–679 (2004)
26. Bodenhausen, G.V., Macrae, C.N., Garst, J.: *Stereotypes in Thought and Deed: Social-Cognitive Origins of Intergroup Discrimination*. Lawrence Erlbaum Associates Publishers, Mahwah (1998)
27. Brewer, M.B.: A dual process model of impression formation. In: *Advances in Social Cognition*, vol. I, pp. 1–36. Psychology Press, London (1988)
28. Bruce, V., Young, A.: Understanding face recognition. *Br. J. Psychol.* **77**(3), 305–327 (1986)
29. Calder, A.J., Young, A.W.: Understanding the recognition of facial identity and facial expression. *Nat. Rev. Neurosci.* **6**(8), 641–651 (2005)
30. Chiao, J.Y., Adams, R.B., Jr, Tse, P.U., Lowenthal, W.T., Richeson, J.A., Ambady, N.: Knowing who's boss: fmri and erp investigations of social dominance perception. *Group Process. Intergr. Relat.* **11**(2), 201–214 (2008)
31. Craig, B.M., Lee, A.J.: Stereotypes and structure in the interaction between facial emotional expression and sex characteristics. *Adapt. Hum. Behav. Physiol.* **6**(2), 212–235 (2020)
32. Craig, B.M., Mallan, K.M., Lipp, O.V.: The effect of poser race on the happy categorization advantage depends on stimulus type, set size, and presentation duration. *Emotion* **12**(6), 1303 (2012)
33. De Gelder, B.: Why bodies? twelve reasons for including bodily expressions in affective neuroscience. *Philos. Trans. R. Soc. B Biol. Sci.* **364**(1535), 3475–3484 (2009)
34. Devine, P.G.: Stereotypes and prejudice: their automatic and controlled components. *J. Pers. Soc. Psychol.* **56**(1), 5 (1989)
35. Driver, J., IV, Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., Baron-Cohen, S.: Gaze perception triggers reflexive visuospatial orienting. *Vis. Cogn.* **6**(5), 509–540 (1999)
36. Dunbar, R.I.: Neocortex size as a constraint on group size in primates. *J. Human Evol.* **22**(6), 469–493 (1992)
37. Dunbar, R.I.: The social brain hypothesis. *Evol. Anthropol. Issues News Rev.* **6**(5), 178–190 (1998)
38. Durik, A.M., Hyde, J.S., Marks, A.C., Roy, A.L., Anaya, D., Schultz, G.: Ethnicity and gender stereotypes of emotion. *Sex Roles* **54**(7), 429–445 (2006)
39. Eagly, A.H., Nater, C., Miller, D.I., Kaufmann, M., Sczesny, S.: Gender stereotypes have changed: a cross-temporal meta-analysis of us public opinion polls from 1946 to 2018. *Am. Psychol.* **75**(3), 301 (2020)

40. Ellsworth, P.C.: Direct gaze as a social stimulus: the example of aggression. *Nonverbal Commun. Aggression* 53–75 (1975)
41. Fehr, B.J., Exline, R.V.: Social Visual Interaction: A Conceptual and Literature Review. Lawrence Erlbaum Associates, Inc, Mahwah (1987)
42. Fiske, S.T., Neuberg, S.L.: A continuum of impression formation, from category-based to individuating processes: influences of information and motivation on attention and interpretation. In: *Advances in Experimental Social Psychology*, vol. 23, pp. 1–74. Elsevier, Amsterdam (1990)
43. Fiske, S.T., Cuddy, A.J., Glick, P., Xu, J.: A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *J. Pers. Soc. Psychol.* **82**(6) (2002)
44. Fox, E., Mathews, A., Calder, A.J., Yiend, J.: Anxiety and sensitivity to gaze direction in emotionally expressive faces. *Emotion* **7**(3), 478 (2007)
45. Frischen, A., Bayliss, A.P., Tipper, S.P.: Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychol. Bull.* **133**(4), 694 (2007)
46. Ganel, T., Goshen-Gottstein, Y., Goodale, M.A.: Interactions between the processing of gaze direction and facial expression. *Vis. Res.* **45**(9), 1191–1200 (2005)
47. Ghavami, N., Peplau, L.A.: An intersectional analysis of gender and ethnic stereotypes: testing three hypotheses. *Psychol. Women Q.* **37**(1), 113–127 (2013)
48. Graham, R., LaBar, K.S.: Garner interference reveals dependencies between emotional expression and gaze in face perception. *Emotion* **7**(2), 296 (2007)
49. Hadjikhani, N., Hoge, R., Snyder, J., de Gelder, B.: Pointing with the eyes: the role of gaze in communicating danger. *Brain Cogn.* **68**(1), 1–8 (2008)
50. Halberstadt, A.G., Castro, V.L., Chu, Q., Lozada, F.T., Sims, C.M.: Preservice teachers' racialized emotion recognition, anger bias, and hostility attributions. *Contemp. Educ. Psychol.* **54**, 125–138 (2018)
51. Halberstadt, A.G., Cooke, A.N., Garner, P.W., Hughes, S.A., Oertwig, D., Neupert, S.D.: Racialized emotion recognition accuracy and anger bias of children's faces. *Emotion* **22**(3), 403 (2022)
52. Haxby, J.V., Hoffman, E.A., Gobbini, M.I.: The distributed human neural system for face perception. *Trends Cogn. Sci.* **4**(6), 223–233 (2000)
53. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**(5539), 2425–2430 (2001)
54. Hess, U., Adams, R., Jr., Kleck, R.E.: Facial appearance, gender, and emotion expression. *Emotion* **4**(4), 378 (2004)
55. Hess, U., Adams, R., Jr., Kleck, R.: Who may frown and who should smile? Dominance, affiliation, and the display of happiness and anger. *Cogn. Emot.* **19**(4), 515–536 (2005)
56. Hess, U., Adams, R., Jr., Kleck, R.E.: Looking at you or looking elsewhere: the influence of head orientation on the signal value of emotional facial expressions. *Motiv. Emot.* **31**, 137–144 (2007)
57. Hess, U., Adams, R., Jr., Grammer, K., Kleck, R.E.: Face gender and emotion expression: are angry women more like men? *J. Vis.* **9**(12), 19–19 (2009)
58. Holmes, E.A., Coughtrey, A.E., Connor, A.: Looking at or through rose-tinted glasses? Imagery perspective and positive mood. *Emotion* **8**(6), 875 (2008)
59. Hori, E., Tazumi, T., Umeno, K., Kamachi, M., Kobayashi, T., Ono, T., Nishijo, H.: Effects of facial expression on shared attention mechanisms. *Physiol. Behav.* **84**(3), 397–405 (2005)
60. Hugenberg, K.: Social categorization and the perception of facial affect: target race moderates the response latency advantage for happy faces. *Emotion* **5**(3), 267 (2005)
61. Hugenberg, K., Bodenhausen, G.V.: Facing prejudice: implicit prejudice and the perception of facial threat. *Psychol. Sci.* **14**(6), 640–643 (2003)
62. Hugenberg, K., Bodenhausen, G.V.: Ambiguity in social categorization: the role of prejudice and facial affect in race categorization. *Psychol. Sci.* **15**(5), 342–345 (2004)

63. Hutchings, P.B., Haddock, G.: Look black in anger: the role of implicit prejudice in the categorization and perceived emotional intensity of racially ambiguous faces. *J. Exp. Soc. Psychol.* **44**(5), 1418–1420 (2008)
64. Johnson, K.L., McKay, L.S., Pollick, F.E.: He throws like a girl (but only when he's sad): emotion affects sex-decoding of biological motion displays. *Cognition* **119**(2), 265–280 (2011)
65. Johnson, K.L., Freeman, J.B., Pauker, K.: Race is gendered: how covarying phenotypes and stereotypes bias sex categorization. *J. Pers. Soc. Psychol.* **102**(1), 116 (2012)
66. Jones, B.C., DeBruine, L.M., Main, J.C., Little, A.C., Welling, L.L., Feinberg, D.R., Tiddeman, B.P.: Facial cues of dominance modulate the short-term gaze-cuing effect in human observers. *Proc. R. Soc. B Biol. Sci.* **277**(1681), 617–624 (2010)
67. Juth, P., Lundqvist, D., Karlsson, A., Öhman, A.: Looking for foes and friends: perceptual and emotional factors when finding a face in the crowd. *Emotion* **5**(4), 379 (2005)
68. Kleinke, C.L.: Gaze and eye contact: a research review. *Psychol. Bull.* **100**(1), 78 (1986)
69. Klucharev, V., Sams, M.: Interaction of gaze direction and facial expressions processing: Erp study. *Neuroreport* **15**(4), 621–625 (2004)
70. Kyriakou, K., Kleanthous, S., Otterbacher, J., Papadopoulos, G.A.: Emotion-based stereotypes in image analysis services. In: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 252–259 (2020)
71. Langton, S.R., Bruce, V.: Reflexive visual orienting in response to the social attention of others. *Vis. Cogn.* **6**(5), 541–567 (1999)
72. Le Gal, P.M., Bruce, V.: Evaluating the independence of sex and expression in judgments of faces. *Percept. psychophys.* **64**(2), 230–243 (2002)
73. Maddox, K.B., Gray, S.A.: Cognitive representations of black americans: reexploring the role of skin tone. *Pers. Soc. Psychol. Bull.* **28**(2), 250–259 (2002)
74. Martinez, L., Falvello, V.B., Aviezer, H., Todorov, A.: Contributions of facial expressions and body language to the rapid perception of dynamic emotions. *Cogn. Emot.* **30**(5), 939–952 (2016)
75. Mathews, A., Fox, E., Yiend, J., Calder, A.: The face of fear: effects of eye gaze and emotion on visual attention. *Vis. Cogn.* **10**(7), 823–835 (2003)
76. Meeren, H.K., van Heijnsbergen, C.C., de Gelder, B.: Rapid perceptual integration of facial expression and emotional body language. *Proc. Natl. Acad. Sci.* **102**(45), 16518–16523 (2005)
77. Milders, M., Hietanen, J.K., Leppänen, J.M., Braun, M.: Detection of emotional faces is modulated by the direction of eye gaze. *Emotion* **11**(6), 1456 (2011)
78. Peterson, J.C., Uddenberg, S., Griffiths, T.L., Todorov, A., Suchow, J.W.: Deep models of superficial face judgments. *Proc. Natl. Acad. Sci.* **119**(17), e2115228119 (2022)
79. Plant, E.A., Hyde, J.S., Keltner, D., Devine, P.G.: The gender stereotyping of emotions. *Psychol. Women Q.* **24**(1), 81–92 (2000)
80. Putman, P., Hermans, E., Van Honk, J.: Anxiety meets fear in perception of dynamic expressive gaze. *Emotion* **6**(1), 94 (2006)
81. Re, D.E., Rule, N.O.: Heavy matters: the relationship between just noticeable differences in perceptions of facial adiposity and facial attractiveness. *Soc. Psychol. Pers. Sci.* **7**(1), 69–76 (2016)
82. Rhue, L.: Racial influence on automated perceptions of emotions. Available at SSRN 3281765 (2018)
83. Richeson, J.A., Todd, A.R., Trawalter, S., Baird, A.A.: Eye-gaze direction modulates race-related amygdala activity. *Group Process. Intergr. Relat.* **11**(2), 233–246 (2008)
84. Robinson, M.D., Johnson, J.T.: Is it emotion or is it stress? Gender stereotypes and the perception of subjective experience. *Sex Roles* **36**(3), 235–258 (1997)
85. Rucker, D.D., Galinsky, A.D., Magee, J.C.: The agentic–communal model of advantage and disadvantage: How inequality produces similarities in the psychology of power, social class, gender, and race. *Adv. Exp. Soc. Psychol.* **58**, 71–125 (2018)
86. Sander, D., Grandjean, D., Kaiser, S., Wehrle, T., Scherer, K.R.: Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion. *Eur. J. Cogn. Psychol.* **19**(3), 470–480 (2007)

87. Sato, W., Yoshikawa, S., Kochiyama, T., Matsumura, M.: The amygdala processes the emotional significance of facial expressions: an fmri investigation using the interaction between expression and face direction. *Neuroimage* **22**(2), 1006–1013 (2004)
88. Senju, A., Hasegawa, T.: Direct gaze captures visuospatial attention. *Vis. Cogn.* **12**(1), 127–144 (2005)
89. Shields, S.A.: Passionate men, emotional women: psychology constructs gender difference in the late 19th century. *Hist. Psychol.* **10**(2), 92 (2007)
90. Tipples, J.: Fear and fearfulness potentiate automatic orienting to eye gaze. *Cogn. Emot.* **20**(2), 309–320 (2006)
91. Trawalter, S., Todd, A.R., Baird, A.A., Richeson, J.A.: Attending to threat: race-based patterns of selective attention. *J. Exp. Soc. Psychol.* **44**(5), 1322–1327 (2008)
92. Vuilleumier, P., George, N., Lister, V., Armony, J., Driver, J.: Effects of perceived mutual gaze and gender on face processing and recognition memory. *Vis. Cogn.* **12**(1), 85–101 (2005)
93. Webbink, P.: The Power of the Eyes. Springer Publishing Co, New York (1986)
94. Weisbuch, M., Adams, R.B., Jr.: The functional forecast model of emotion expression processing. *Soc. Pers. Psychol. Compass* **6**(7), 499–514 (2012)
95. Wilkins, C.L., Chan, J.F., Kaiser, C.R.: Racial stereotypes and interracial attraction: phenotypic prototypicality and perceived attractiveness of asians. *Cult. Divers. Ethn. Minor. Psychol.* **17**(4), 427 (2011)
96. Wilson, J.P., Hugenberg, K., Rule, N.O.: Racial bias in judgments of physical size and formidability: From size to threat. *J. Pers. Soc. Psychol.* **113**(1), 59 (2017)
97. Zebrowitz, L.A., Kikuchi, M., Fellous, J.M.: Facial resemblance to emotions: group differences, impression effects, and race stereotypes. *J. Pers. Soc. Psychol.* **98**(2), 175 (2010)

Chapter 4

Social Vision of the Body in Motion: Interactions Between the Perceiver and the Perceived



Pamala N. Dayley and Kerri L. Johnson

Abstract The human body reliably conveys meaningful information to observers. Some factors originate in the target of perception. These include static aspects of body shape and also dynamic aspects of body motion. The perception of bodies in motion also relies on factors that originate in the perceiver, including extant knowledge structures (e.g., stereotypes and prior experiences) that influence the decoding of phenotypic cues. This chapter will review three distinct questions. First, how the visible cues of targets of perception and a perceiver's prior knowledge interact and bias the perception of dynamic body cues in the target, specifically as they relate to emotion perception. Second, how such patterns also exert a meaningful influence on the performance of identities, with a specific emphasis on how gendered body motions impact judgments of sexual orientation. Finally, we review evidence of how targets of perception strategically alter their movements to engender desirable perceptions in observers, with an emphasis on how people strategically alter their gait to appear more attractive to observers. Collectively, these findings highlight the importance of understanding both how cues convey information to observers and how observers' prior experiences and knowledge structures can bias their perceptions.

4.1 Introduction

Well, you can tell by the way I use my walk, I'm a woman's man... —Bee Gees

P. N. Dayley

Department of Psychology, University of California, Los Angeles, CA, USA

e-mail: pnd5035@g.ucla.edu

K. L. Johnson (✉)

Department of Communication, University of California, Los Angeles, CA, USA

Department of Psychology, University of California, Los Angeles, CA, USA

e-mail: kerri.johnson@comm.ucla.edu

The second single released from the Saturday Night Fever soundtrack, Stayin' Alive aptly describes the scene depicted in the opening credits. John Travolta embodies the character of an unabashedly straight man swaggering down a busy New York City sidewalk. His movements draw the visual attention of passersby, likely intentionally, and undoubtedly evoke reliable interpersonal inferences from those observers. Here and throughout the film, the body's movement is central to cinematic art. Although many aspects of the film might be considered cringe-worthy by modern standards, the centrality of the human body in motion for interpersonal perceptions rings true to this day, both artistically and scientifically. Indeed, scientific discoveries confirm that the human body in motion provides potent cues that communicate meaningful information to observers. Here, we describe such evidence.

In everyday social interactions, humans rely on the visual characteristics of others to form impressions and judgments about them [13]. Much of the existing literature has probed how the human face conveys meaningful visual information to others. This is perhaps unsurprising given the multifaceted cues and information value involved in face perception, including physical structure [18, 28, 52], dynamic expressions [11, 17], and even eye gaze [1, 50]. And yet, the face is not the only way that observers ascertain information about unknown others. Visual information provided by the human body is equally as rich as that provided by faces. Indeed, both face and body perception involve similar cognitive mechanisms such as holistic processing [16], similar developmental trajectories [47], and even the integration of both static and dynamic information [17]. And yet, relative to faces, bodies have the potential to provide observers with unique information that could not otherwise be obtained. As such, body perception could (and perhaps should) be regarded as equally important in social perception. Indeed, several noteworthy aspects of human bodies distinguish it from faces and their perception.

One factor that distinguishes the perception of faces versus bodies is physical size. Not surprisingly, bodies occupy greater structural mass than faces. This enables two unique aspects of body perception. First, body cues can be discerned at greater physical distances that would otherwise compromise accurate face perception [10]. Second, many body cues are largely viewpoint independent, allowing observers to process information from visual vantages that preclude face perception (e.g., a target's body can be seen from behind but a target's face cannot).

A second factor that distinguishes body perception from face perception is that body motions have the potential to carry more meaningful social information than faces (see e.g., [3]). Observers form impressions of others based on cues from both the face and body. Knowing this, targets of perception are prone to use their face and body to communicate to observers, thus eliciting more favorable impressions. However, these efforts are asymmetric, such that people tend to monitor and control their facial expressions more extensively than their body motions thus allowing their true states to "leak" more from body than facial cues. Consequently, compared to judgments of facial expressions, judgments of bodies tend to be more accurate along a range of dimensions [2]. Relatedly, judgments of bodies can also inform an appropriate response in a given context. Whereas faces are

sufficient to communicate an appropriate emotional reaction (Be fearful!), bodies also communicate reasonable actions to take (Run!).

Arguably, perceptions of the body can be even more potent than those of the face, particularly when they involve interactions between a target of perception and a perceiver. Both dynamic cues such as the body's velocity and trajectory as well as static cues such as the body's shape and size provide independent sources of visual information to observers. Additionally, perceivers rely on existing knowledge, preferences, and prejudices to form impressions of others. Together, information existing within the target of perception and the perceiver influence the social perception process.

4.2 Determinants of Body Perception Originating in the Target of Perception Dynamic Cues

4.2.1 *Dynamic Cues*

Much of the earlier perceptual research on body perception sought to determine whether and how human observers perceived the dynamic motion of human bodies. Gunnar Johansson leveraged techniques that had been used within the arts, but infrequently by scientists (see e.g., [30]). He invited research participants to his laboratory where he mounted illuminated bulbs to each person's major joints. In a darkened room, Johansson filmed his participants as they engaged in a range of physical actions. The resulting films depicted only illuminated points of light, which later came to be known as point-light displays (PLDs). Subsequently, naïve observers viewed and provided judgments about these films, and the findings were illuminating. Not only were these impoverished PLDs sufficient to support the perception that the displays portrayed human movement (i.e., animacy), they were also sufficient to inform accurate impressions of the type of movement being enacted (e.g., running versus walking, jumping jacks, etc.) [22]. This foundational insight inspired an entire subfield of perceptual science to investigate the perception of biological motions. Although beyond the scope of this brief review, a few key findings are worth noting.

Dynamic body motions inform a range of categorical social judgments. Once created, PLDs are shown to observers who judge them on a range of social dimensions, and they achieve a high degree of accuracy given the minimal visual information that is presented. For example, observers can distinguish whether a target identifies as male or female [8, 9, 20, 23, 42] largely because of sexually dimorphic body motion trajectories [4, 9, 32]. Other social categorizations also occur when participants view PLDs, including sexual orientation [25, 29], race [35], age [14, 36, 37], threat/strength [15, 33], and emotion [6, 10, 26, 41]. Importantly, we note that much (but not all) of the existing literature has yet to incorporate more

diverse gender and sexual identities, either as participants or targets of perception, although we are optimistic that this is actively changing (see e.g., [39]).

Dynamic body motions also inform more evaluative social judgments. Indeed, both PLDs and dynamic avatars have been shown to observers who evaluated their attractiveness. Johnson et al. [25] presented dynamic silhouettes that varied in gendered body motions (from a masculine shoulder swagger to a more feminine hip sway). Perhaps unsurprisingly they found that observers preferred masculine walk motions when targets were judged to be men, but feminine motions when targets were judged to be women. Thus, the body's dynamic movement is sufficient to inform both categorical and evaluative social judgments.

4.2.2 Structural Cues

Static and structural aspects of the body, such as its shape and size, are also potent cues that convey meaningful information to observers. As is true for the perception of dynamic body motions, differences in body shape inform both categorical and evaluative social judgments.

In his seminal work, Devendra Singh [48] provided early insights regarding the potency of body shape for evaluative social judgments. He developed a set of static stimuli in which he systematically manipulated body shape of line-drawn female targets, resulting in stimuli that varied in the circumference ratio of the waist to the hips (i.e., the waist-to-hip ratio (WHR)). He presented these drawings to observers who judged each woman for attractiveness and fecundity. Observers judged women with low WHRs to be more attractive and fecund than women with high WHRs. He reasoned that the WHR provides an honest indicator of women's health and fertility, and as such, it is a key factor in judgments of attractiveness [48]. Although these patterns replicated in Western samples (see e.g., [49]), they were less robust and even opposite in other cultural contexts that lacked exposure to Western media [5, 31, 39].

Interestingly, the very factor that determined attractiveness judgments in Singh's early work is implicated in other social judgments as well. The WHR, a cue that is anthropometrically sexually dimorphic, is sufficient to inform sex category judgments [23, 27]. It is also used to infer a target's sexual attitudes, with low WHR women being judged as more sexually unrestricted, relative to higher WHR women [40].

The impact of body shape in social judgments also extends beyond the WHR. The presence of adipose tissue (i.e., body fat), for instance, is linked to negative social evaluations [21], generally, and interpersonal discrimination [7] and dehumanization [45], specifically. That said, heavier bodies are also judged to pose less threat to observers [46].

Thus, body perception, although less extensively explored scientifically, proves to be as potent as faces for social perception. With this foundation in mind, we now turn to more nuanced findings within the body perception literature that contextual-

izes how visible body cues are perceived. Such research is highly interdisciplinary, involving scholars from multiple allied social and cognitive sciences, and it has implications for computational approaches to decoding body cues.

4.3 Integration of Visible Cues

Above we reviewed how aspects of both dynamic and static body cues inform social judgments. Essentially, we highlighted how cues that originate in the target of perception inform social judgments. Yet, these cues are not perceived in a vacuum. Instead, human perceivers tend to integrate all available information both within and between sensory modalities. Consequently, the reliable impact of both dynamic and static body cues on social perception can be accentuated or attenuated.

Efforts to isolate body motion through the use of PLDs or body shape through the use of carefully controlled line drawings allows scholars to isolate the independent impact of a particular cue on social judgments. In the wild, however, these cues are rarely considered in isolation, but rather in combination with other cues. Numerous empirical studies have documented how multiple visible cues contextualize one another's perception.

For instance, although we argued above that body perception can, at times, be more important than face perception because it can be achieved at physical distances or visual vantages that preclude face perception, that is not always the case. We routinely encounter individuals for whom both the face and body are visible. In such circumstances, the independent information provided by the body appears to impact the perception of the face, and vice versa [51] (see also [34], cf. [43]). In one study, participants viewed an image in which the emotion being expressed by the face and body were either congruent or incongruent, and they matched the emotion portrayed by the body to one of two exemplars, one of which matched the body expression and one that did not. Overall, participants were faster to identify the matching body emotion when the facial expression had been congruent, rather than incongruent, with the body expression [51]. In another study, participants viewed images in which facial expressions had been morphed from fear to anger. These were superimposed over bodies that expressed either fear or anger. Although participants were instructed to judge facial expressions and ignore body expressions, their responses indicated that this did not occur. Instead, judgments of facial emotion were consistently influenced by body expressions [51].

Similar integration of visible cues occurs when observers integrate information from body shape and body motion during their simultaneous perception. As noted above, body shape and motion are independently sufficient to support sex categorizations, yet they are more routinely perceived simultaneously. In such instances, body shape is prone to inform categorical judgments of sex (i.e., male/female), and body shape is prone to inform more continuous judgments of gender (i.e., masculine to feminine; [23]). When sex categorizations rely strictly on body motion, observers appear to infer sex categories from more continuous gendered impressions such that

body motions that are perceived to be masculine are judged to be male, and body motions that are perceived to be feminine are judged to be female [23].

Similar effects occur for other social judgments, including evaluations of perceived attractiveness and categorizations of sexual orientation. The perceived attractiveness of body motion only occurs after sex categorization is achieved [24]. Thus, observers first use the body's shape to determine a target's sex category membership. Thereafter, body shape and motion combinations are deemed attractive to the extent that they convey compatible information: female-shaped bodies moving in feminine ways and male-shaped bodies moving in masculine ways. Incompatible combinations (i.e., combinations that are perceived as feminine men or masculine women) tend to be judged as less attractive. These simple patterns of compatibility also predict perceptions of sexual orientation, such that when shape and motion cues are congruent, targets tend to be judged as straight, but when shape and motion cues are incongruent, targets tend to be judged as gay [27].

4.4 Perceiver Influences

In addition to multiple visible cues contextualizing one another's perception, perceivers themselves also influence the process based on prior experiences, emotions, and knowledge structures.

4.4.1 *Self-protective Biases*

The experience of various emotions, for example, is sufficient to shift social categorizations for body cues. One simple example of this occurs when people are making sex categorizations from body shapes. Although body shape is sexually dimorphic, the distribution of phenotypes for men and women overlap. Such overlaps yield some degree of ambiguity when observers are making social judgments. Particularly under such conditions of uncertainty, observers are prone to shift the threshold for making male versus female categorizations. For example, in one set of studies, participants categorized the sex of body silhouettes that varied systematically in WHR [23]. Their judgments showed an overall pattern that indicated a male-categorization default, which is typical of social judgments more generally. Yet the threshold at which WHRs shifted from being judged as female to being judged as male shifted depending on the emotion being experienced by participants such that the threshold was relaxed somewhat when participants were induced to feel joy, relative to fear, in part because men tend to be perceived as more threatening and formidable than women.

In other research, similar “self-protective biases” have been shown to shift the perception of race. In one set of studies, participants categorized the race of PLDs. Under threat, PLDs were more likely to be categorized as Black, rather than

White [35]. Thus, one's own emotional experience has the potential to influence the social perception of body cues.

4.4.2 Knowledge Structures and Stereotypes

In other instances, prior knowledge structures, or stereotypes, are known to bias social judgments of bodies. As noted above, sex categorizations of bodies tend to occur spontaneously, integrating information from both body shape and motion. When only body motion is visible, observers infer sex category membership from perceived masculinity/femininity [23]. This inferential process allows other gender-linked stereotypes to exert a similar influence on the perception of PLDs. Arm motion depicted in PLDs, for example, informs a variety of social judgments, including emotional affect. Indeed, observers can distinguish basic emotions from arm motions ranging from knocking to throwing a ball [41]. Because emotions are highly gender-stereotyped (see [19] for face perception equivalent), sex categorizations are influenced by the emotion being depicted. In a series of studies [26], participants categorized PLD arm motions to be either male or female. Although the emotion being conveyed in each PLD varied orthogonally to the target's actual sex, it nevertheless influenced judgments in a way that was consistent with gender stereotypes. When viewing PLDs expressing anger, perceivers were more likely to categorize the target as a man, and when viewing PLDs expressing sadness, perceivers were more likely to categorize the target as a woman.

4.5 Deliberate Manipulation of Own Body Movement

The social perception of bodies can be modulated by factors that originate in the target of perception (e.g., the simultaneous perception of multiple cues) and in the perceiver (e.g., emotion states or prior knowledge structures/stereotypes). In addition to those factors, it is also important to acknowledge that targets of perception themselves can deliberately modify their body's motion to engender desired percepts in others.

Because perceivers rely on gendered body motion to evaluate a target and the outcomes of such evaluations are predictable, targets of perception can alter gendered aspects of their gait (i.e., an individual's manner of walking) to evoke a particular social identity categorization from others [44], (see also [12]). Two specific judgments that may be altered include evaluations of sexual orientation and attractiveness. In one study that tested walk motion differences as a function of sexual orientation [29], gay and straight-identified research participants were induced to walk with either gender-typical or gender-atypical gaits. These recordings were converted to PLDs that were shown to a naïve group of observers. Based on these PLDs, perceivers' accuracy in categorizing each target's sex showed an

intriguing pattern of results. First, the sex categorizations were more accurate for judgments of straight-identified than for gay-identified men and women overall. Yet this pattern differed when targets of perception had been instructed to enact gender-atypical walk motions. Then, although sex categorizations of straight-identified targets remained well above chance, the opposite was true for gay-identified targets. Specifically, when gay-identified targets were induced to move in a gender-atypical manner, the resulting PLDs reliably compelled a cross-sex categorization among observers. The authors argue that these individuals might have been prone to more closely monitor their motions to shift the perceptions of others, at times aiming to conceal their sexual identity and at times aiming to convey their sexual identity to others.

Additionally, because gendered gaits influence perceptions of attractiveness, they also have the potential to be harnessed to enhance one's attractiveness to others. In recent work [44] (see also [12]), targets of perception were instructed to walk in specific ways while their motions were being recorded. In addition to walking naturally, in separate trials these individuals were also instructed to move in a gender-typical manner and in a way that would enhance their physical attractiveness to observers. The authors hypothesized that because gender-typical movements tend to be judged as more attractive, in general, individuals who aim to appear attractive might do so by enhancing the gender typicality of their gait [15]. Indeed, the findings were consistent with this hypothesis. Attractiveness judgments tended to be more favorable when targets of perception aimed to appear both gender-typical and attractive, relative to their natural walk motion. This overall pattern of results suggests that human body movement is subject to deliberate modification that can enhance evaluative judgments made by observers. In other work, sartorial decisions made by women in particular, shift one's center of gravity and body movements in a manner that also enhances perceived attractiveness among observers [38]. Thus, deliberate changes to motion or incidental changes to one's footwear reliably alter the percepts among observers.

An additional way that people might alter their body's motion has a more serious impact than social evaluations. Instead, they determine one's physical safety. Gunns et al. [15] noted that perpetrators of violent physical attacks report that they select their victims based on the perceived ease of attack, based in part on their gait. The authors recruited two groups of individuals. One group underwent self-defense training that would render them less vulnerable to attack. The other group engaged in walk motion training to mitigate the perception of vulnerability. PLDs of the members of each group were evaluated by naïve observers. Changes to perceived vulnerability were more robust following walk motion training than following self-defense training. This important insight provides actionable steps¹ that anyone can enact to reduce their perceived vulnerability. Although the authors provide a highly scientific analysis of the specific joint angles that reduce perceived vulnerability, the overall pattern can be described simply to novice readers: walk like a man! Walk

¹ Apologies for the gratuitous academic pun.

motion training that taught people to walk with their head high, their chest out, and with a lateral (confident) walk motion proved most effective in mitigating perceived vulnerability.

4.6 Conclusion

As reviewed in this chapter, the existing body perception literature identifies four primary takeaway concepts. First, while body perception and face perception are rooted in the same basic visual process, they have unique facets that deserve dedicated attention separate from each other. The primary factors that distinguish perceptions of bodies from perceptions of faces include physical size and emotion regulation. These factors allow for unique evaluations of the target that are less practical for facial perception. Second, there exists a two-part relationship between the target and perceiver in which both parties provide information that aids in informing impressions of the target. Targets provide visual indicators including dynamic cues (e.g., walk motion) and structural cues (i.e., body shape, body size, WHR) that perceivers use to make later judgments about the target. However, perceivers also rely on their own emotions, beliefs, and knowledge to further inform judgments about the target. Third, body motion exerts a meaningful and stable influence on the evaluation and performance of identities. Specifically, evaluations of a target's sex, attractiveness, and sexual orientation are reliably rated among perceivers. As detailed previously, certain combinations of body shape and body motion increase evaluations of attractiveness (e.g., congruent body cues (female-shaped bodies moving in feminine ways) are rated as more attractive) and evoke unique evaluations of sexual orientation (e.g., congruent body cues (female-shaped bodies moving in feminine ways) are more commonly judged as straight while incongruent body cues (female-shaped bodies moving in masculine ways) are more often judged as gay). Finally, conveying, exaggerating, or concealing specific body motions enables targets of perception to lead perceivers to identify them using certain categorizations. Specifically, targets may alter their body motion to alter evaluations of their sexual orientation, increase their perceived attractiveness, and decrease their perceived vulnerability. Future research in the field should aim to build upon this information by looking at other evaluations that may be altered through deceptive body motion as well as body motion in groups. Taking a multi-disciplinary perspective that includes individuals from social and cognitive sciences as well as from the arts, researchers will be better equipped to investigate the perception of body movement in future research.

References

1. Adams Jr, R.B., Kleck, R.E.: Perceived gaze direction and the processing of facial displays of emotion. *Psychol. Sci.* **14**(6), 644–647 (2003)
2. Ambady, N., Rosenthal, R.: Nonverbal communication. *Encyclopedia Mental Health* **2**, 775–782 (1998)
3. Aviezer, H., Trope, Y., Todorov, A.: Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* **338**(6111), 1225–1229 (2012)
4. Barclay, C.D., Cutting, J.E., Kozlowski, L.T.: Temporal and spatial factors in gait perception that influence gender recognition. *Percept. Psychophys.* **23**, 145–152 (1978)
5. Bovet, J.: Evolutionary theories and men's preferences for women's waist-to-hip ratio: Which hypotheses remain? a systematic review. *Front. Psychol.* **10**, 1221 (2019)
6. Chouchourelou, A., Matsuoka, T., Harber, K., Shiffra, M.: The visual analysis of emotional actions. *Soc. Neurosci.* **1**(1), 63–74 (2006)
7. Crandall, C.S., D'Anello, S., Sakalli, N., Lazarus, E., Nejatardt, G.W., Feather, N.: An attribution-value model of prejudice: Anti-fat attitudes in six nations. *Personal. Soc. Psychol. Bull.* **27**(1), 30–37 (2001)
8. Cutting, J.E.: Generation of synthetic male and female walkers through manipulation of a biomechanical invariant. *Perception* **7**(4), 393–405 (1978)
9. Cutting, J.E., Proffitt, D.R., Kozlowski, L.T.: A biomechanical invariant for gait perception. *J. Exp. Psychol. Hum. Percept. Perform.* **4**(3), 357 (1978)
10. De Gelder, B.: Towards the neurobiology of emotional body language. *Nat. Rev. Neurosci.* **7**(3), 242–249 (2006)
11. Dobs, K., Bülothoff, I., Schultz, J.: Use and usefulness of dynamic face stimuli for face perception studies—a review of behavioral findings and methodology. *Front. Psychol.* **9**, 1355 (2018)
12. Fink, B., Weege, B., Neave, N., Pham, M.N., Shackelford, T.K.: Integrating body movement into attractiveness research. *Front. Psychol.* **6**, 220 (2015)
13. Freeman, J.B., Johnson, K.L.: More than meets the eye: Split-second social perception. *Trends Cogn. Sci.* **20**(5), 362–374 (2016)
14. Galusca, C.I., Quinn, P.C., Heron-Delaney, M., Pascalis, O.: Infant sensitivity to age-based social categories in full-body displays. *Infant Behav. Dev.* **68**, 101726 (2022)
15. Gunns, R.E., Johnston, L., Hudson, S.M.: Victim selection and kinematics: A point-light investigation of vulnerability to attack. *J. Nonverbal Behav.* **26**(3), 129–158 (2002)
16. Harris, A., Vyas, D.B., Reed, C.L.: Holistic processing for bodies and body parts: New evidence from stereoscopic depth manipulations. *Psychon. Bull. Rev.* **23**, 1513–1519 (2016)
17. Hehman, E., Flake, J.K., Freeman, J.B.: Static and dynamic facial cues differentially affect the consistency of social evaluations. *Personal. Soc. Psychol. Bull.* **41**(8), 1123–1134 (2015)
18. Hehman, E., Leitner, J.B., Deegan, M.P., Gaertner, S.L.: Picking teams: When dominant facial structure is preferred. *J. Exp. Soc. Psychol.* **59**, 51–59 (2015)
19. Hess, U., Adams Jr, R.B., Grammer, K., Kleck, R.E.: Face gender and emotion expression: Are angry women more like men? *J. Vis.* **9**(12), 19–19 (2009)
20. Hiris, E., Conway, S., McLoughlin, W., Yang, G.: Individual observer differences in the use of form and motion to perceive the actor's sex in biological motion displays. *Percept. Motor Skills* **129**(1), 5–32 (2022)
21. Hu, Y., Parde, C.J., Hill, M.Q., Mahmood, N., O'Toole, A.J.: First impressions of personality traits from body shapes. *Psychol. Sci.* **29**(12), 1969–1983 (2018)
22. Johansson, G.: Visual perception of biological motion and a model for its analysis. *Percept. Psychophys.* **14**, 201–211 (1973)
23. Johnson, K.L., Tassinary, L.G.: Perceiving sex directly and indirectly: Meaning in motion and morphology. *Psychol. Sci.* **16**(11), 890–897 (2005)
24. Johnson, K.L., Tassinary, L.G.: Compatibility of basic social perceptions determines perceived attractiveness. *Proc. Natl. Acad. Sci.* **104**(12), 5246–5251 (2007)

25. Johnson, K.L., Gill, S., Reichman, V., Tassinary, L.G.: Swagger, sway, and sexuality: Judging sexual orientation from body motion and morphology. *J. Personal. Soc. Psychol.* **93**(3), 321 (2007)
26. Johnson, K.L., McKay, L.S., Pollick, F.E.: He throws like a girl (but only when he's sad): Emotion affects sex-decoding of biological motion displays. *Cognition* **119**(2), 265–280 (2011)
27. Johnson, K.L., Iida, M., Tassinary, L.G.: Person (mis) perception: Functionally biased sex categorization of bodies. *Proc. R. Soc. B Biol. Sci.* **279**(1749), 4982–4989 (2012)
28. Joshi, M.P., Lloyd, E.P., Diekman, A.B., Hugenberg, K.: In the face of opportunities: Facial structures of scientists shape expectations of stem environments. *Personal. Soc. Psychol. Bull.*, 01461672221077801 (2022)
29. Lick, D.J., Johnson, K.L., Gill, S.V.: Deliberate changes to gendered body motion influence basic social perceptions. *Soc. Cogn.* **31**(6), 656–671 (2013)
30. Marey, E.J.: Movement. Arno Press and New York Times, New York (1895/1982)
31. Marlowe, F., Wetsman, A.: Preferred waist-to-hip ratio and ecology. *Personal. Individual Differences* **30**(3), 481–489 (2001)
32. Mather, G., Murdoch, L.: Gender discrimination in biological motion displays based on dynamic cues. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* **258**(1353), 273–279 (1994)
33. McCarty, K., Hönekopp, J., Neave, N., Caplan, N., Fink, B.: Male body movements as possible cues to physical strength: a biomechanical analysis. *Am. J. Hum. Biol.* **25**(3), 307–312 (2013)
34. Meeren, H.K., van Heijnsbergen, C.C., de Gelder, B.: Rapid perceptual integration of facial expression and emotional body language. *Proc. Natl. Acad. Sci.* **102**(45), 16518–16523 (2005)
35. Miller, S.L., Maner, J.K., Becker, D.V.: Self-protective biases in group categorization: Threat cues shape the psychological boundary between “us” and “them”. *J. Personal. Soc. Psychol.* **99**(1), 62 (2010)
36. Montepare, J.M., Zebrowitz, L.A.: Person perception comes of age: The salience and significance of age in social judgments. In: *Advances in Experimental Social Psychology*, vol. 30, pp. 93–161. Elsevier (1998)
37. Montepare, J.M., Zebrowitz-McArthur, L.: Impressions of people created by age-related qualities of their gaits. *J. Personal. Soc. Psychol.* **55**(4), 547 (1988)
38. Morris, P.H., White, J., Morrison, E.R., Fisher, K.: High heels as supernormal stimuli: How wearing high heels affects judgements of female attractiveness. *Evol. Hum. Behav.* **34**(3), 176–181 (2013)
39. Oswald, F., Adams Jr, R.B.: Feminist social vision: Seeing through the lens of marginalized perceivers. *Personal. Soc. Psychol. Rev.*, 10888683221126582 (2022)
40. Paganini, G.A., McConnell, A.A., Deska, J.C., Almaraz, S.M., Hugenberg, K., Lloyd, E.P.: Waist-to-hip ratio predicts sexual perception and responses to sexual assault disclosures. *Personal. Soc. Psychol. Bull.*, 01461672221148008 (2023)
41. Pollick, F.E., Paterson, H.M., Bruderlin, A., Sanford, A.J.: Perceiving affect from arm movement. *Cognition* **82**(2), B51–B61 (2001)
42. Pollick, F.E., Kay, J.W., Heim, K., Stringer, R.: Gender recognition from point-light walkers. *J. Exp. Psychol. Hum. Percept. Perform.* **31**(6), 1247 (2005)
43. Reed, C.L., Bukach, C.M., Garber, M., McIntosh, D.N.: It's not all about the face: Variability reveals asymmetric obligatory processing of faces and bodies in whole-body contexts. *Perception* **47**(6), 626–646 (2018)
44. Shropshire, J., Johnson, K.L.: Something in the way she (or he) moves: The role of walk motion in sex categorization and attractiveness ratings. Manuscript submitted for publication (2023)
45. Sim, M., Almaraz, S.M., Hugenberg, K.: Bodies and minds: Heavier weight targets are de-mentalized as lacking in mental agency. *Personal. Soc. Psychol. Bull.* **48**(9), 1367–1381 (2022)
46. Sim, M., Almaraz, S.M., Hugenberg, K.: Stereotyping at the intersection of race and weight: diluted threat stereotyping of obese black men. *J. Exp. Soc. Psychol.* **99**, 104274 (2022)
47. Simion, F., Di Giorgio, E., Leo, I., Bardi, L.: The processing of social stimuli in early infancy: from faces to biological motion perception. *Progr. Brain Res.* **189**, 173–193 (2011)
48. Singh, D.: Adaptive significance of female physical attractiveness: role of waist-to-hip ratio. *J. Personal. Soc. Psychol.* **65**(2), 293 (1993)

49. Singh, D., Dixson, B.J., Jessop, T.S., Morgan, B., Dixson, A.F.: Cross-cultural consensus for waist-hip ratio and women's attractiveness. *Evol. Hum. Behav.* **31**(3), 176–181 (2010)
50. Steiner, T., Brandenburg, J., Adams Jr, R.B.: The effects of facial dominance and gender prototypicality on the gaze-cuing effect. *J. Vis.* **16**(12), 1398–1398 (2016)
51. Van den Stock, J., Righart, R., De Gelder, B.: Body expressions influence recognition of emotions in the face and voice. *Emotion* **7**(3), 487 (2007)
52. Wilson, J.P., Hugenberg, K., Rule, N.O.: Racial bias in judgments of physical size and formidability: From size to threat. *J. Personal. Soc. Psychol.* **113**(1), 59 (2017)

Chapter 5

Visual Perception of Threat: Structure, Dynamics, and Individual Differences



Kestutis Kveraga

Abstract Efficient recognition of potential threats greatly enhances an organism's chances of survival. This involves the ability to rapidly identify and react to clear, imminent danger, balanced by a more deliberative process when assessing ambiguous threat situations. While many models of affective perception posit that all negative stimuli are threatening and aversive, and thus should be treated similarly, our studies show that human observers keenly discriminate different types of negative stimuli, exhibiting distinct behavioral and neural response patterns. We have shown that humans rapidly extract the spatial and temporal qualities of threat from scene images to decipher the structure of the threat—whether a clear threat is present, the likelihood of physical harm, how imminent it is, and its spatial direction—whether it is aimed at the observer or someone else. These qualities can be quickly extracted from full-color scenes as well as from simple line drawings. Our findings in psychophysical and imaging studies examining the three major visual streams—the magnocellular, parvocellular, and koniocellular pathways—suggest that they may contribute to different phases of the threat perception process. Lastly, we will delve into how characteristics such as anxiety, sex, and hemispheric laterality affect behavioral and neural responses in perceiving threat.

5.1 Introduction

Visual perception of threat is the front end of a complex process whose ultimate goal is to maximize the safety and survival chances of the perceiver. As vision normally is humans' keenest, most dominant sense, we rely on it to continuously analyze an incoming stream of information to scan for danger. The process must be finely balanced, as missing threats can have grave consequences, but overreacting to stimuli that turn out to be innocuous (i.e. false alarms) is also counterproductive.

K. Kveraga (✉)

Beth Israel Deaconess Medical Center and Athinoula A. Martinos Center for Biomedical Imaging
at Mass General Brigham, Harvard Medical School, Boston, MA, USA
e-mail: kestas@nmr.harvard.edu

Visual threats can be very clear, such as someone pointing a gun in your face, or quite subtle, such as minuscule changes in the facial expression, body posture, or movements of someone about to attack you. The facial and bodily expressions of someone about to attack you do not have to overtly signal anger, and in real life situations, unlike in laboratory studies, they typically do not. Slight narrowing of the eyes and flaring of the nostrils, tension and lowering of the jaw, partially turning (“blading”) the body with the dominant side on the back foot, reducing the interpersonal distance, and having the dominant hand by the side of the body can alert you that the person is loading up on a punch or a strike with a weapon. And sometimes it is the absence, rather than the presence, of something—for example, the lack of usual activity and presence of people around you, or a lack of traffic on the road, that can alert those with relevant experience to grave danger, such as a roadside bomb or an ambush nearby.

It would seem that recognizing threat is a most elementary function that most cognitively intact adults should perform effortlessly. We are all progeny of those before us who survived and passed on their genes because they were good at recognizing and avoiding threats. However, in relatively safe Western societies, this ability seems somewhat eroded or suppressed, as some individuals fall prey to violent crime partly because they seem unable identify potential predators and threat situations and take preventive measures. The flip side of recognizing threat is recognizing the lack of threat projected by potential victims. Interviews with convicted violent criminals indicate that they tend to pick their victims not at random, but from a subset of people who walk around seemingly unaware of the threat posed by potential victimizers nearby and unsure of themselves, while avoiding those who appear to be mindful of their surroundings and confident. Violent criminals can pick up these cues from gestures and movements of the potential victims within seconds, with lack of attention to the surroundings, asynchronous limb movements, abnormal gait, and slumping posture being prime predictors of being selected as a victim [18, 37]. While individuals with frequent exposure to danger, such as frontline combat troops and law enforcement personnel, are well attuned to the relevant threat cues due to their experience and training, a scientific and broader public understanding of how threat stimuli are processed by the mind and brain to distinguish them from other negative stimuli, is lacking.

The first, automatic determination the brain makes upon exposure to a stimulus is its affective valence—whether the stimulus is potentially good or bad for the organism [12], with negative stimuli having a greater impact than positive stimuli [15]. While threat stimuli tend to be negative, they are not always purely negative (as discussed below). Conversely, there are many kinds of negative stimuli that are not threatening (e.g., those evoking sadness, disgust, disappointment, frustration and the like). Despite this, all leading models of affective perception are two-dimensional, defining stimuli as some combination of qualities corresponding to how positive or negative, and how engaging they are [25]. These two dimensions are usually called valence and arousal, or something akin to it [13, 14]: misery-pleasure and sleep-arousal [69]; unpleasantness-pleasantness and disengagement/engagement [79]; tension-calmness and tiredness-energy [75]; unpleasant-pleasant and low activation-

high activation [58]. Moreover, different types of unpleasant stimuli, such as those depicting threat, gore, injury, death, defeat, disgust, sadness and pity, all can evoke varied levels of negative valence and arousal and thus cannot be distinguished by their coordinates in the valence-arousal plane. Because there is no unique dimension which separates the negative threatening and non-threatening stimuli [55], these two-dimensional models of emotion tend to equate threat and non-threat negative stimuli, while both our experience and research suggests that we react to threat stimuli differently than to other negative stimuli.

5.2 How Do We Distinguish Threat from Other Negative Stimuli?

Previous research using threat stimuli by Ohman et al. [65] has shown that certain facial expressions, such as facial anger, and natural threats, such as spiders and snakes, are identified faster than non-threatening stimuli like mushrooms and flowers. However, it is not clear what specific mechanisms are responsible for perceiving these stimuli as threatening rather than benign.

Given the lack of clarity about how our brains recognize threat objects in visual images, we wanted to identify the dimensions that distinguish visual threat stimuli from other negative, non-threat stimuli, as well as attempting to understand the qualities that make some threat stimuli different from other threat stimuli. Some researchers have suggested that the ancestors of hominids, the earliest small primates, had evolved a threat response based on visual patterns and movements of snakes, their natural predators [47] and, supporting that hypothesis, neurons sensitive to images of snakes have been found in the pulvinar nucleus of the thalamus of neonate monkeys without previous exposure to snakes [77]. However, humans are also able to quickly identify manmade threat objects which are too recent to have brain templates shaped by evolution, such as guns, edged weapons and sharp tools [17]. It thus appears that, with sufficient exposure, humans are able to quickly learn object shapes that present threat in modern environments.

Aside from predatory animals and natural disasters, other humans are a major source of threat to us, particularly when living in modern and urban environments. Humans, perhaps more than any other animal because of our exposed facial skin and white sclera, signal their intent and internal state via facial cues, such as facial expression, color, and eye gaze [1, 3, 35, 57, 70]. While facial expressions of emotion can be a useful threat cue, in real life facial expressions are often ambiguous, seemingly inappropriate for the context, or even deceptive. Humans are also quite capable of masking their true feelings of anger by keeping their facial expression impassive or even smiling while getting ready to attack. Thus, facial expressions, while sometimes useful, can be poor indicators of an impending threat from the expresser. Other, no less important cues are usually present in determining whether someone presents a threat, and have been somewhat neglected in emotion research.

We will discuss these threat cues and how they comprise the major dimensions of threat next.

5.3 Threat Dimensions

In our research examining how humans distinguish threat from other negative stimuli, we have identified visual cues that, in combination, create the structure of threat based on three dimension. The dimensions of this structure are those signaling threat *direction, imminence and capacity* to do harm.

5.3.1 Threat Direction

Threat direction is conveyed by the position, orientation, and movement cues of the eye gaze, head, body, and limbs of the potential attacker. For example, if the eye gaze, face and body of someone in a confrontation are directed towards the perceiver, either squarely or at a slight angle, with the dominant side of the body turned at an angle (“bladed”) in a split stance and dominant hand clenched by the side, this often signals an impending attack by loading up on a punch or using a weapon.

Once the person in a confrontation with you has initiated the attack, the body turns to produce the power behind the punch or thrust with a cutting/striking weapon, or to aim and align a firearm with the dominant eye. The timing of the threat has changed from imminent to happening right now, but the spatial focus of the threat has not. The most salient spatial characteristic of direct threat is the focus of the attack being directed towards the observer.

5.3.2 Threat Imminence

Threat imminence refers to the timing of the act intended to cause harm to you or someone else. Timing is determined from the status of the victim and the presence (or absence) and actions of the threat agent. If the threat agent appears to be getting ready to attack, or the attack is already in progress, it constitutes direct or indirect threat scene, depending on the orientation of the attack. If the threat agent is absent and the victim appears to be injured or dead, or the threat appears to be already executed, it constitutes what we term as a non-threat negative, or “Threat Aftermath” scene. Such images tend to attract greater attention and exploration (see below) but slower responses than current threat images. The reasons for this may have to do with the greater complexity of the images, seeking to extract knowledge about the causes and consequences of the executed threat from threat aftermath scenes.

5.3.3 Threat Capacity

The third critical dimension of threat is the ability to inflict harm. In humans, this is conveyed by many relatively stable cues signaling biological sex, age, strength, hormone levels, as well as more fluid cues, such as those indicating mental state. In the face, this is expressed as the masculinity or maturity cues [81]—a prominent brow, deep-set eyes, strong jaw, facial hirsuteness and skin hue, which is usually darker for males than females for every racial group [70]. By definition, the maturity cues also signal age, an important factor in the ability to inflict harm, as the very young and very old have less threat capacity and inclination to aggression than the so-called fighting-age males. Strength cues are implicitly represented in the facial structure by the prominence of the brow, nose and jaw, and by the face width/height ratio (FWHR), with relatively broader faces associated with greater strength, aggression, dominance, inclination to psychological and physical threat, and levels of circulating testosterone in males [59, 60, 80]. Moreover, FWHM was shown to directly predict athletic success in professional athletes [24, 76], and fighting capacity (“formidability”) in professional fighters [82]. In combination, compound visual cues signaling masculinity (such as facial shape, hue and hirsuteness), strength and fighting ability (FWHM) and age interplay to convey threat capacity.

The other cues in humans are body shape and posture, and weaponry. Facial and bodily cues and convey strength and therefore threat capacity independently and jointly, via interaction [7, 34]. Weapons, of course, both increase the danger of serious or fatal damage, and can change the equation of natural threat capacity favoring the armed person. In non-human threats, large predators that can overpower humans to cause injury or death, either singly or in packs, or smaller predators that cause incapacitation via venomous strikes or stings, are deemed to have the highest threat capacity.

5.3.4 Evidence for the Three Threat Dimensions

In our initial studies of how humans distinguish threat from other negative images, we collected over 500 negative and context-matching neutral color images. In analyzing the images, it became apparent that while some images depicted impending threat to the observer, others showed threat or harm being done to others, while yet others suggested the harm had already occurred, with no impending danger at present. In the fourth category, images had similar threat context (e.g., a weapon or a dangerous predator), but the threat was well controlled. To test these perceptions, we sorted the images into what we termed Direct Threat, Indirect Threat, Merely Negative/Threat Aftermath and Neutral categories. If an image depicted an imminent or in-progress attack angled towards the observer, we termed such images Direct Threat scenes [55] (Fig. 5.1 left-most panels). If the spatial focus of the impending or in-progress attack was directed away from the observer,



Fig. 5.1 (a) Examples of color photographic images with varying amounts of threat. (b) Examples of the line drawing scenes made from the same color photographs [19, 55]

towards another person (or animal), we termed images in this category Indirect Threat scenes. Lastly, images which depicted or implied harm suffered by people or animals, without an active threat evident in the image, we called Merely Negative or Threat Aftermath scenes. Note that the images in all three categories were negative in affective valence. We then asked subjects to rate them on three questions that we had determined were most relevant to verify these categories (Harm to You, Harm to Others, Past Harm; see Fig. 5.2 and [55] for details). The last category were Low Threat images used as “neutral” control stimuli in the same general context.

Subjects were able to keenly discriminate the Direct Threat, Indirect Threat, and Threat Aftermath/Merely Negative images, producing distinct behavioral rating patterns for each category. We then scanned a different group of subjects using fMRI while they performed a 1-back working memory task on a selected subset of the images from each category. We found that Direct Threat, Indirect Threat,

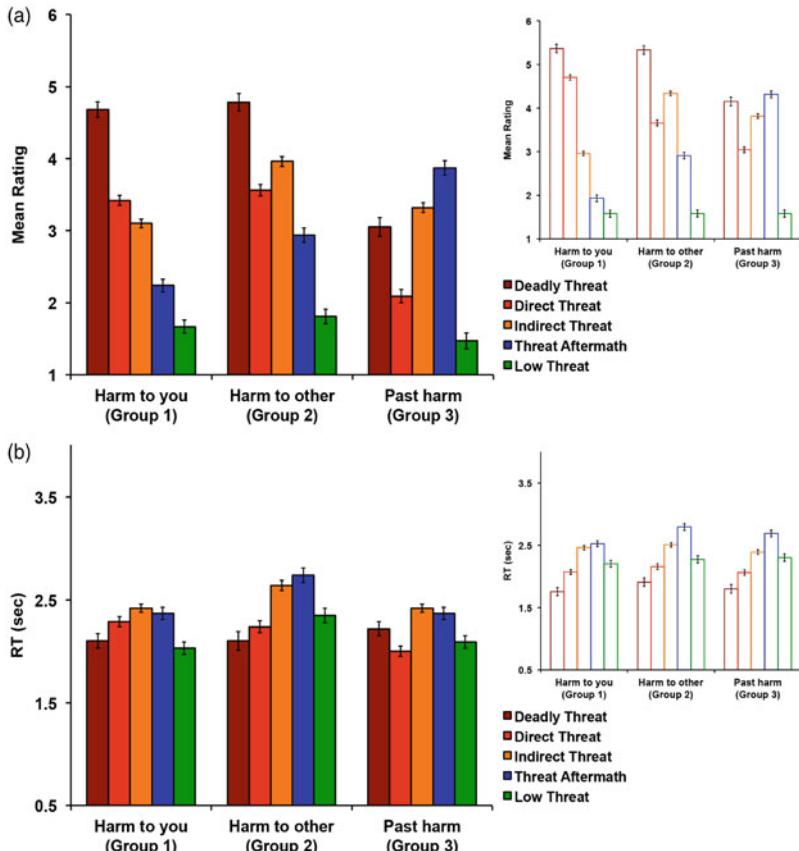


Fig. 5.2 (a) Ratings for color photograph (solid bars) and line-drawing (outline bars) images. In reddish tones are the ratings for the deadly (brown), direct (brick), and indirect (orange) threat images, blue bars are the negative non-threat images, and green bars are the “neutral”, context-matched, non-threat, non-negative images. (b) Mean response times to rate these images. Adapted from [19]

and Merely Negative images produce distinct, albeit partly overlapping, pattern of activations while participant viewed the images and performed a working memory task (Fig. 5.2). The Direct Threat images activated the periaqueductal gray, the amygdala, and the ventral and dorsal prefrontal cortex more, while the Threat Aftermath images engaged the medial prefrontal, parahippocampal, and posterior cingulate cortex to a greater extent [55]. This latter pattern of activations may have emerged because of greater processing of contextual associations engaged by these complex scene images [8, 9, 54].

In visual discrimination of threat and negative images, images in certain categories may contain associated salient cues (e.g., exposed teeth in predatory animals, blood), which could have affected the subjects’ responses. We thus wanted to

test these categories with stimuli that were highly similar in their textural and color content. In a follow-up study by [19], we created and tested 500 simplified monochromatic line drawings made from the original color images, asking subjects to rate the stimuli on the same three harm dimensions as in [55]. The ratings and responses times for the line drawing scenes were highly similar to the ratings found for the original color images (Fig. 5.2). A cluster analysis of both the color image and line drawing scenes rating also identified a subcategory of Direct Threat comprising especially dangerous threats we termed the Deadly Threat images. This subcategory included mostly humans threatening with deadly weapons or in-progress attacks by large animal predators [19]. From the findings of this study, we concluded that human observers are able to quickly and efficiently identify threats of different types from very basic features captured in the monochromatic line drawings.

5.4 The Role of Context in Threat Perception

To discriminate innocuous stimuli from those that should be treated as a threat, we rely not just on analysis of visual facial, bodily, movement and object cues, but also on the context in which they occur. Context is one of the key top-down influences in the brain trying to predict its environment [53, 64] and is extremely important not just in threat perception, but in all vision, with out-of-context objects recognized less efficiently. In potential threat situations, the time to make a response decision is compressed and the cost of errors greater than during “normal” perception. Therefore, under conditions of uncertainty and time pressure, we may operate with incomplete (not fully analyzed) sensory information and rely more on the context to make decisions. The context during threat perception is useful in assigning probabilities to particular events—imagine getting up at night without turning on the lights and glimpsing a dark shape in the hallway. If you live in a very safe neighborhood and share your home with family members who tend to be up at night, it is probably no cause for alarm. It would be ill-advised to overreact in a situation where your household members’ being up at night is not at all unusual. But if you live alone, in a neighborhood where break-ins are not uncommon, you should be greatly alarmed and immediately take defensive measures. Thus, context sets boundaries on what is essentially an error-detection process, with error in this case defined as a deviation from the expected state of affairs for that contextual setting.

While context is usually helpful in making the correct response decision, in rare cases it can also lead the responder astray. A police officer who assumes the presence (or absence) of a gun in an encounter with a civilian is potentially making life-and-death decisions. Research using first-person-shooter paradigms using civilians and police officers suggests that stereotypic influences set a context that can lead to misidentifying a benign object as a gun or vice versa and lead to an erroneous decision to shoot, or not to shoot [27, 28, 74].

In reference to the images we tested in [19], context, along with the facial and bodily cues, plays a role in whether the situation is considered threatening. While the alligator and the gun in Fig. 5.1 are most dangerous when oriented towards the observer “in the wild” (as in the leftmost column), an alligator behind plate glass in a zoo or a gun used in target practice at a shooting range should evoke considerably less alarm, as the threat ratings for images in this category suggest (Fig. 5.2, top right panel, green bars). Observers rated such as images as low threat, because while the object (alligator or a loaded gun) can still be dangerous, it is controlled in those particular contexts.

5.5 Hemispheric Lateralization in Threat Perception

The brains of humans and many other animals have long been known to have developmental asymmetries and functional specialization, dating back to [20]. As concerns perception and responding to threat, the majority of humans are right-hand dominant [62] and perceive threat stimuli faster in the left visual hemifield (which projects to the right hemisphere; [68]). Therefore, it is usually the left side of the potential attacker that is turned closer to the perceiver, which affords better threat perception from the left visual hemifield, projected to the right hemisphere. This would be reversed for a left hand-dominant attacker and, because of the predominance of right-handedness, confers a fighting advantage to left-handed fighters due to the surprise and less experience defending attacks coming from the right side of the observer (e.g., the increased incidence of left-handedness in warrior tribes, in boxing and mixed martial arts [67]). While right-handedness (“pawedness”) is less pronounced with animals, they, too, tend to exhibit the left-side bias in attacking, presumably because of the right-hemisphere advantage in processing threat stimuli (see [68], for a review).

To investigate hemispheric lateralization during perception of threat and negative stimuli, we presented either a threat or non-threat negative image paired with a context-matched neutral (non-threat, non-negative) image. The images were presented simultaneously and bilaterally, with subjects asked to keep their gaze on a centrally presented fixation cross at the start of each trial. Subjects had to identify which of two bilaterally presented images contained a threat or negative scene via a left- or right-hand key press aligned with the image, but otherwise were free to view the images freely to make their response decision after the fixation cross was removed. The latency of key press responses was shorter when a threat or negative image appeared on the left, rather than on the right, and much shorter for threat vs. non-threat negative images (Fig. 5.3b). While participants were not asked to make a saccade to either of the stimuli, we recorded their eye movements while they viewed the bilateral scenes and made a response decision. The saccade latency was shorter when the image on the left contained a threat rather than a negative scene, but the opposite was true for the initial saccades to images presented on the right (Fig. 5.3a). Lastly, subjects made more saccades and dwelled longer when negative

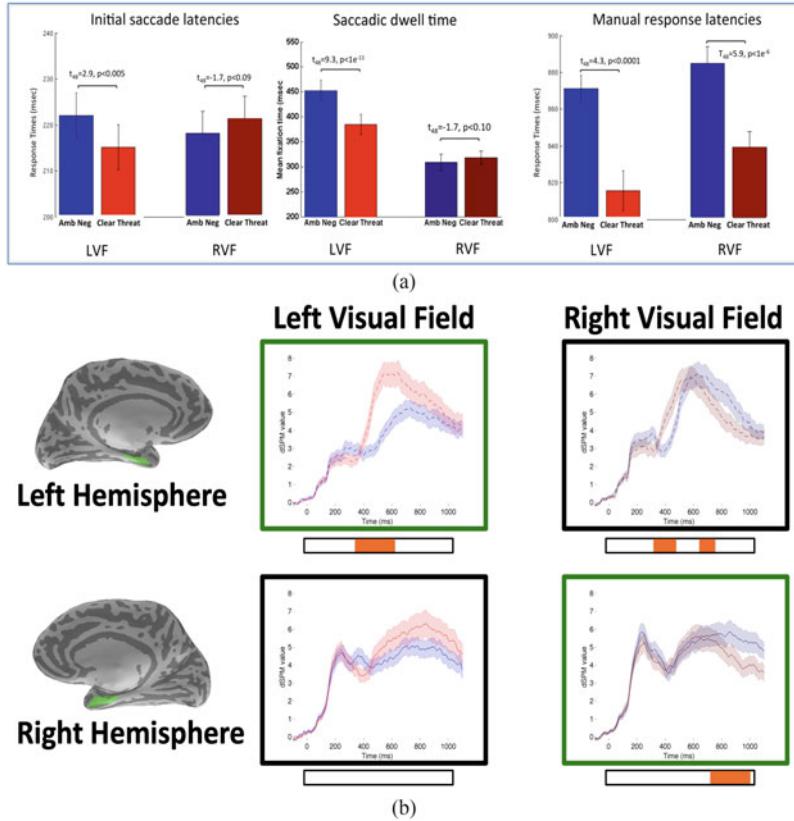


Fig. 5.3 (a) Saccadic and manual responses to Direct Threat and Threat Aftermath scenes presented in the left or right visual hemifield. (b) MEG activations in the periamygdaloid cortex to the same stimuli. Red-hued bars and lines are responses to Direct Threat stimuli, responses to Ambiguous Negative Scenes are in blue hues [78]

and threat images were presented on the left, and spent more time exploring the negative non-threat images rather than the clear threat images. These findings seem to confirm a proclivity of humans to respond faster to threat images appearing in the left hemifield ([78], Fig. 5.3a). A subsequent MEG study using the same paradigm found that Direct Threat scenes presented to the left visual hemifield evoked larger and earlier peaks of activity in the periamygdaloid cortex, while Threat Aftermath scenes evoked greater late activity ([78], Fig. 5.3b).

5.6 Clear and Ambiguous Threat in Faces

In perceiving threat from faces, facial expression is combined with other threat cues, such as facial maturity, strength cues, as discussed in the Threat Capacity section. Another important facial cue is eye gaze. Humans are capable of reading mental state of others just from the eye region [4]. The direction of eye gaze also contextualizes facially expressed emotion and threat. For example, an anger expression coupled with a direct gaze presents a clear threat stimulus in that the person is expressing anger at the observer. This makes it a direct threat, while anger combined with an averted eye gaze implies the threat is directed at another person and weakens the threat. Conversely, a fear expression with an averted gaze clearly points to the source of threat, while direct gaze combined with a fear expression renders the source of threat ambiguous. Separable routes have been proposed to mediate the discrimination of eye gaze and facial expression [21], but studies by Adams et al. [1, 3, 5] showed that the eye gaze cues strongly interact with facial expression. In the following studies we used this paradigm to test the hypothesis that the magnocellular (M) pathway will be preferentially attuned to clear threat cue combinations (e.g., averted-gaze fear or direct-gaze anger), while the parvocellular (P) pathway will be more attuned to threat ambiguity (direct-gaze fear or averted-gaze anger). We showed that this was indeed the case in a number of fMRI and MEG studies [6, 30, 31].

Here we investigated how observers' trait anxiety modulates M and P pathway processing of clear and ambiguous threat cues. Again, clear threat cues are combinations of cues that together clearly signal threat, whereas ambiguous threat combinations arise when the cues contradict one another or are neutral [2]. Faces convey several dynamic and stable cues, such as facial expression and eye gaze (both dynamic), age/maturity and masculinity (relatively stable or slow-changing), and biological sex, race, ethnicity, and social group or class (stable). Focusing on the dynamic face cues, Adams et al. have shown that facial anger combined with a direct gaze towards the observer signal clear threat, whereas an angry face with an averted gaze present ambiguous threat signals ("whom is this person angry towards?") [1, 3, 5]. This is so because anger expressions are an approach cue, while an averted eye gaze is an avoidance cue. Therefore, a face expressing fear (an avoidance cue) combined with an averted gaze present clear threat cues, but fear combined with direct gaze is ambiguous ("Is the person afraid of me? Seeking help?"). Anger and direct gaze both signal approach by the source of the threat, whereas fear and averted gaze both signal avoidance of threat.

5.7 The Role of the Major Visual Pathways in Threat Perception

The three major pathways in the human visual system—the magnocellular (M), parvocellular (P), and koniocellular (K) subdivisions—arise in the retinal ganglion cells that combine the inputs of rods and cones of different types, and project through the lateral geniculate nucleus of the thalamus, continuing to the primary visual cortex (V1) and beyond. The M pathway originates in the large “parasol” ganglion cells of the retina, summing the inputs of many long- and medium-wavelength cones, as well as rods, whereas parvocellular (P) pathway originates in the midget retinal ganglion cells, which typically differentiate inputs from just two red-green bipolar cells connected to a single cone (see [53] for a review). The M cells are prevalent in the retinal periphery, have large receptive fields, are highly sensitive to luminance contrast and motion, but are not color-sensitive or able to resolve fine details. Conversely, the P cells have small receptive fields, are red-green sensitive, and are less responsive to small luminance changes and high-frequency motion. The M projections comprise much of the dorsal, or “where”, visual stream, which specializes in global spatial vision, motion detection, attention, and action planning whereas the P projections form much of the ventral visual (“what”) stream [36, 63].

The third, koniocellular (K) pathway, is formed from the speck-like (“konio”) retinal ganglion cells that combine the inputs of short-wavelength sensitive (S) cones with the combined inputs of long and medium wavelength cones and has qualities that are somewhat complementary to those of the M and P pathways [26, 44]. The K pathway is thought to be a phylogenetically older pathway that serves as the color channel in the many species of mammals with dichromatic vision. Because of these qualities, it may be particularly attuned to identifying biologically relevant stimuli, including threat stimuli. Indeed, Isbell et al. [47, 77] have proposed that the koniocellular pathway is of particular evolutionary importance, in that its processing qualities are best suited for detecting the primary predators of the early mammals—snakes. The “Snake Detection Theory” [47] received some support when neurons sensitive to snakes were found in neonate macaques without prior exposures to snakes [77]. Interestingly, the stimuli with second-most responses were macaque faces with emotional expressions. A subcortical “low road” snaking through the superior colliculus, pulvinar nucleus, and the amygdala was proposed as a low-latency pathway using K pathway inputs for detecting predatory threats such as snakes [47].

Whether the shortest-latency saccades to threat stimuli are triggered via K inputs, and which threat cues the K pathway might be attuned to in humans, was not known. The superior colliculus is necessary for generating the very low-latency saccades (known as “express saccades”) and until recently was thought to be responsive primarily to achromatic stimuli (implying M-pathway inputs). However, recent evidence from electrophysiology studies by Hall and Colby [41, 42] in the macaque

used carefully calibrated violet-blue geometric stimuli to show that neurons in the SC respond strongly to K-biased inputs.

To investigate whether the K or another pathway may be involved in fast orienting responses to threat stimuli, we conducted an eye-tracking study in humans. As threat stimuli, we used angry and neutral faces presented laterally, and biased to preferentially engage the M, P, or K pathways [56]. Subjects made saccades to the face stimuli with the lowest latency to the K-biased faces, which also evoked the greatest proportion of express saccades, the low-latency saccades that form a distinct subpopulation of all refixation saccades. This finding suggests that the K pathway may be involved in the early, orienting phase of threat perception, perhaps via a pathway that projects from the superior colliculus to the pulvinar nucleus of the thalamus and then the amygdala. A recent imaging study in humans using high-resolution, high-field (7 Tesla) fMRI and path mediation analyses found support for such a pathway with aversive stimuli, though it did not test which of the three (M, P, or K) visual inputs might contribute to it [50].

The M pathway bias towards fast processing of coarse stimuli may also make the M pathway well suited for processing clear threat cues. Bar [10] proposed that the M pathway, because of its fast, coarse projections to the dorsal stream and the prefrontal cortex, may play a role in rapidly activating prototypical category templates, which then provide top-down guidance and facilitation to slower, more detailed visual processing taking place along the ventral temporal visual processing regions. For example, a coarse, low spatial frequency (LSF) image of a mushroom may activate templates associated with a mushroom or an umbrella, constraining the possibilities of what the object might be and biasing processing towards the most likely possibilities. This supposition was born out in several studies (e.g., [11, 52]) which showed that low spatial frequency (LSF) and M-biased stimuli indeed activate the ventral prefrontal and orbitofrontal cortex. The former study, using MEG, showed that this activation in OFC precedes the activation in the fusiform, object processing, region of the ventral temporal lobe [11]. The latter study used dynamic causal modeling to show that activity in OFC exerts top-down influence on the object processing regions in the fusiform gyrus of the ventral temporal lobe [52].

Furthermore, we found that objects presented to the M pathway were recognized faster than objects presented to the P pathway, despite the latter being rated by subjects as being more visible [52]. In several imaging studies using clear and ambiguous threat cues, in faces and in scene images, we then showed that the M pathway preferentially processes visual stimuli projecting clear threat (e.g., [6, 30, 31, 45]). This was tested with face images that expressed fear, in which eye gaze was manipulated to be direct or averted. Recall that a fearful face with an averted gaze is a clear threat cue, because indicates a clear source of threat from the direction of the gaze [3], while a fear face with a direct gaze (looking at you) is ambiguous, because it is unclear where the source of threat is—whether it is you, behind you, or whether the person showing fear is seeking help. In studies using fMRI and MEG we manipulated the face images to be biased towards the M or P pathway, by individual adjustment of luminance and color contrast, such

that the M images were grayscale and had low luminance contrast (<5%), and thus invisible to the P pathway, or red/green isoluminant and therefore invisible to the M pathway. Using fMRI and MEG, we repeatedly found that the M-pathway was more engaged by clear threat cue combinations (averted-gaze fear), while the P pathway showed more engagement by ambiguous threat cues [6, 31, 45, 46]. These findings are consistent with the processing characteristics of the M and P pathways—the fast, coarse projection to the dorsal visual stream subserving action for the M pathway, and the more deliberate, detailed processing in the ventral visual stream devoted to visual form analysis [53].

5.8 Individual Differences

5.8.1 Threat and Anxiety

In a study using facial expressions crossed with averted or direct eye gaze, we investigated how observers' trait anxiety influences M and P pathway processing of clear and ambiguous threat cues. Our subjects ($N=108$) exhibiting a wide range of trait anxiety were scanned with fMRI while they viewed fearful or neutral faces with averted or directed gaze. The luminance and color of face stimuli were calibrated to selectively engage the M- or P-pathways. We found that higher trait anxiety facilitated processing of clear threat projected to M-pathway, but resulted in impaired perception of ambiguous threat projected to P-pathway. Increased right amygdala reactivity was associated with higher anxiety for M-biased clear threat cues (averted-gaze fear), while increased left amygdala reactivity was associated with higher anxiety for P-biased ambiguous threat cues (direct gaze crossed with a fear expression). This lateralization became more pronounced with higher levels of trait anxiety. Our findings suggest that trait anxiety enhances perception of clear threat (averted-gaze facial fear) via increased right amygdala activity and with increased left amygdala reactivity to ambiguous (direct-gaze fear) facial threat cues [45].

5.8.2 Sex Differences in Threat Processing

Sex differences are known to influence a number of human brain and behavioral functions, including linguistic [72], navigational [38], defensive [48], mathematical [40], and attentional skills [61]. Importantly for threat perception, men and women substantially differ in processing of affective stimuli [71, 73]. Females tend to be more expressive [51] and reactive than males [16], and show stronger

psychophysiological responses to emotional stimuli [66]. However, sex differences in brain responses to *threat* stimuli have not been extensively characterized. Given the critical role of the amygdala in facial expression and eye gaze perception [5, 39, 45], this study sought to examine the functional and anatomical differences in the amygdalae of female vs. male observers. Our subjects engaged in perception of emotional faces biased to engage the M and P pathways and containing threat cues extracted from combined emotional expression and eye gaze direction cues. We focused on the activation of the left and right amygdalae, because they play a critical role in the processing of affective information in general [29, 49], and in threat vigilance specifically [32]. It was previously found that in male observers' brains, the right amygdala was more engaged, while in female observers' brains, the left amygdala was more involved in affective processing [22, 23, 43].

We used two facial expressions (fearful and neutral) and two eye gaze directions (direct and averted) as in many previous studies on facial threat cue perception, and asked observers to identify emotion of the face (e.g., [1, 3, 5, 33, 45]). In addition, we manipulated the luminance and color contrast of face stimuli to selectively engage M or P pathways during facial threat perception. We found that female observers showed more accurate behavioral responses to faces with averted gaze and greater left amygdala reactivity both to fearful and neutral faces. Conversely, males showed greater right amygdala activation only for M-biased averted-gaze fear faces (i.e., clear threat faces). In addition to functional differences, in females bilateral amygdala volumes were larger when adjusted for brain size, and positively correlated with behavioral accuracy for M-biased clear threat faces. In male observers, it was only the right amygdala volume that was positively correlated with accuracy for M-biased clear threat faces. These findings suggest that male observers tend to be more right-lateralized in terms of the amygdala activity and volume, at least when it comes to processing clear threat cues in the face [46].

5.9 Summary

Here we described the dimensions by which threat stimuli are distinguished from other negative stimuli and categorized and results from behavioral, eye-tracking, fMRI and MEG studies characterizing responses to threat and merely negative stimuli. Using facial threat stimuli, we showed that the major visual projections—the K, M, and P pathways—may contribute to different aspects of the threat perception process: orienting to threat (K), conveying clear threat (M), and deliberative responses to ambiguous threat (P). We also described how perceivers' trait anxiety, sex-linked brain differences and hemispheric lateralization influence threat perception.

References

1. Adams Jr, R.B., Kleck, R.E.: Perceived gaze direction and the processing of facial displays of emotion. *Psychol. Sci.* **14**(6), 644–647 (2003)
2. Adams Jr, R.B., Kveraga, K.: Social vision: Functional forecasting and the integration of compound social cues. *Rev. Philos. Psychol.* **6**, 591–610 (2015)
3. Adams Jr, R.B., Gordon, H.L., Baird, A.A., Ambady, N., Kleck, R.E.: Effects of gaze on amygdala sensitivity to anger and fear faces. *Science* **300**(5625), 1536–1536 (2003)
4. Adams Jr, R.B., Rule, N.O., Franklin Jr, R.G., Wang, E., Stevenson, M.T., Yoshikawa, S., Nomura, M., Sato, W., Kveraga, K., Ambady, N.: Cross-cultural reading the mind in the eyes: an fmri investigation. *Journal of Cognitive Neuroscience* **22**(1), 97–108 (2010)
5. Adams Jr, R.B., Franklin Jr, R.G., Kveraga, K., Ambady, N., Kleck, R.E., Whalen, P.J., Hadjikhani, N., Nelson, A.J.: Amygdala responses to averted vs direct gaze fear vary as a function of presentation speed. *Soc. Cogn. Affect. Neurosci.* **7**(5), 568–577 (2012)
6. Adams Jr, R.B., Im, H.Y., Cushing, C., Boshyan, J., Ward, N., Albohn, D.N., Kveraga, K.: Differential magnocellular versus parvocellular pathway contributions to the combinatorial processing of facial threat. *Progr. Brain Res.* **247**, 71–87 (2019)
7. Albohn, D.N., Brandenburg, J.C., Kveraga, K., Adams Jr, R.B.: The shared signal hypothesis: Facial and bodily expressions of emotion mutually inform one another. *Attention, Perception, & Psychophysics* **84**(7), 2271–2280 (2022)
8. Aminoff, E.M., Kveraga, K., Bar, M.: The role of the parahippocampal cortex in cognition. *Trends Cogn. Sci.* **17**(8), 379–390 (2013)
9. Bar, M.: A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* **15**(4), 600–609 (2003)
10. Bar, M., Aminoff, E.: Cortical analysis of visual context. *Neuron* **38**(2), 347–358 (2003)
11. Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Dale, A.M., Hämäläinen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., et al.: Top-down facilitation of visual recognition. *Proc. Natl. Acad. Sci.* **103**(2), 449–454 (2006)
12. Barrett, L.F.: Valence is a basic building block of emotional life. *J. Res. Personal.* **40**(1), 35–55 (2006)
13. Barrett, L.F., Bliss-Moreau, E.: Affect as a psychological primitive. *Adv. Exp. Soc. Psychol.* **41**, 167–218 (2009)
14. Barrett, L.F., Russell, J.A.: The structure of current affect: Controversies and emerging consensus. *Curr. Directions Psychol. Sci.* **8**(1), 10–14 (1999)
15. Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D.: Bad is stronger than good. *Rev. Gen. Psychol.* **5**(4), 323–370 (2001)
16. Birnbaum, D.W., Croll, W.L.: The etiology of children's stereotypes about sex differences in emotionality. *Sex Roles* **10**, 677 (1984)
17. Blanchette, I.: Snakes, spiders, guns, and syringes: How specific are evolutionary constraints on the detection of threatening stimuli? *Q. J. Exp. Psychol.* **59**(8), 1484–1504 (2006)
18. Book, A., Costello, K., Camilleri, J.A.: Psychopathy and victim selection: The use of gait as a cue to vulnerability. *J. Interpersonal Viol.* **28**(11), 2368–2383 (2013)
19. Boshyan, J., Feldman Barrett, L., Betz, N., Adams Jr, R.B., Kveraga, K.: Line-drawn scenes provide sufficient information for discrimination of threat and mere negativity. *i-Perception* **9**(1), 2041669518755806 (2018)
20. Berker, E.A., Berker, A.H., Smith, A.: Translation of Broca's 1865 report: Localization of speech in the third left frontal convolution. *Arch. Neurol.* **43**(10), 1065–1072 (1986)
21. Bruce, V., Young, A.: Understanding face recognition. *Br. J. Psychol.* **77**(3), 305–327 (1986)
22. Cahill, L., Haier, R.J., White, N.S., Fallon, J., Kilpatrick, L., Lawrence, C., Potkin, S.G., Alkire, M.T.: Sex-related difference in amygdala activity during emotionally influenced memory storage. *Neurobiol. Learn. Memory* **75**(1), 1–9 (2001)

23. Canli, T., Zhao, Z., Brewer, J., Gabrieli, J.D., Cahill, L.: Event-related activation in the human amygdala associates with later memory for individual emotional experience. *J. Neurosci.* **20**(19), RC99 (2000)
24. Carré, J.M., McCormick, C.M.: In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proc. R. Soc. B Biol. Sci.* **275**(1651), 2651–2656 (2008)
25. Carroll, J.M., Yik, M.S., Russell, J.A., Barrett, L.F.: On the psychometric principles of affect. *Rev. Gen. Psychol.* **3**(1), 14–22 (1999)
26. Casagrande, V.: A third parallel visual pathway to primate area v1. *Trends Neurosci.* **17**(7), 305–310 (1994)
27. Correll, J., Park, B., Judd, C.M., Wittenbrink, B.: The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals. *J. Personal. Soc. Psychol.* **83**(6), 1314 (2002)
28. Correll, J., Park, B., Judd, C.M., Wittenbrink, B., Sadler, M.S., Keesee, T.: Across the thin blue line: police officers and racial bias in the decision to shoot. *J. Personal. Soc. Psychol.* **92**(6), 1006 (2007)
29. Costafreda, S.G., Brammer, M.J., David, A.S., Fu, C.H.: Predictors of amygdala activation during the processing of emotional stimuli: a meta-analysis of 385 PET and fMRI studies. *Brain Res. Rev.* **58**(1), 57–70 (2008)
30. Cushing, C.A., Im, H.Y., Adams Jr, R.B., Ward, N., Albohn, D.N., Steiner, T.G., Kveraga, K.: Neurodynamics and connectivity during facial fear perception: The role of threat exposure and signal congruity. *Sci. Rep.* **8**(1), 2776 (2018)
31. Cushing, C.A., Im, H.Y., Adams Jr, R.B., Ward, N., Kveraga, K.: Magnocellular and parvocellular pathway contributions to facial threat cue processing. *Soc. Cogn. Affect. Neurosci.* **14**(2), 151–162 (2019)
32. Davis, M., Whalen, P.J.: The amygdala: vigilance and emotion. *Mol. Psychiatry* **6**(1), 13–34 (2001)
33. Ewbank, M.P., Fox, E., Calder, A.: The interaction between gaze and facial expression in the amygdala and extended amygdala is modulated by anxiety. *Front. Hum. Neurosci.*, 56 (2010)
34. de Gelder, B., Van den Stock, J.: Real faces, real emotions: perceiving facial expressions in naturalistic contexts of voices, bodies, and scenes. In: Oxford Handbook of Face Perception. Oxford University Press (2011)
35. Gibson, J.J., Pick, A.D.: Perception of another person's looking behavior. *Am. J. Psychol.* **76**(3), 386–394 (1963)
36. Goodale, M.A., Milner, A.D.: Separate visual pathways for perception and action. *Trends Neurosci.* **15**(1), 20–25 (1992)
37. Grayson, B., Stein, M.I.: Attracting assault: Victims' nonverbal cues. *J. Commun.* **31**(1), 68–75 (1981)
38. Grön, G., Wunderlich, A.P., Spitzer, M., Tomczak, R., Riepe, M.W.: Brain activation during human navigation: gender-different neural networks as substrate of performance. *Nat. Neurosci.* **3**(4), 404–408 (2000)
39. Hadjikhani, N., Hoge, R., Snyder, J., de Gelder, B.: Pointing with the eyes: the role of gaze in communicating danger. *Brain Cogn.* **68**(1), 1–8 (2008)
40. Haier, R.J., Benbow, C.P.: Sex differences and lateralization in temporal lobe glucose metabolism during mathematical reasoning. *Dev. Neuropsychol.* **11**(4), 405–414 (1995)
41. Hall, N., Colby, C.: S-cone visual stimuli activate superior colliculus neurons in old world monkeys: Implications for understanding blindsight. *J. Cogn. Neurosci.* **26**(6), 1234–1256 (2014)
42. Hall, N.J., Colby, C.L.: Express saccades and superior colliculus responses are sensitive to short-wavelength cone contrast. *Proc. Natl. Acad. Sci.* **113**(24), 6743–6748 (2016)
43. Hamann, S.B., Ely, T.D., Grafton, S.T., Kilts, C.D.: Amygdala activity related to enhanced memory for pleasant and aversive stimuli. *Nat. Neurosci.* **2**(3), 289–293 (1999)
44. Hendry, S.H., Reid, R.C.: The koniocellular pathway in primate vision. *Annu. Rev. Neurosci.* **23**(1), 127–153 (2000)

45. Im, H.Y., Adams Jr, R.B., Boshyan, J., Ward, N., Cushing, C.A., Kveraga, K.: Observer's anxiety facilitates magnocellular processing of clear facial threat cues, but impairs parvocellular processing of ambiguous facial threat cues. *Sci. Rep.* **7**(1), 15151 (2017)
46. Im, H.Y., Adams Jr, R.B., Cushing, C.A., Boshyan, J., Ward, N., Kveraga, K.: Sex-related differences in behavioral and amygdalar responses to compound facial threat cues. *Hum. Brain Map.* **39**(7), 2725–2741 (2018)
47. Isbell, L.A.: Snakes as agents of evolutionary change in primate brains. *J. Hum. Evol.* **51**(1), 1–35 (2006)
48. Kline, J.P., Allen, J.J., Schwartz, G.E.: Is left frontal brain activation in defensiveness gender specific? *J. Abnormal Psychol.* **107**(1), 149 (1998)
49. Kober, H., Barrett, L.F., Joseph, J., Bliss-Moreau, E., Lindquist, K., Wager, T.D.: Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *Neuroimage* **42**(2), 998–1031 (2008)
50. Kragel, P.A., Čeko, M., Theriault, J., Chen, D., Satpute, A.B., Wald, L.W., Lindquist, M.A., Barrett, L.F., Wager, T.D.: A human colliculus-pulvinar-amygdala pathway encodes negative emotion. *Neuron* **109**(15), 2404–2412 (2021)
51. Kring, A.M., Gordon, A.H.: Sex differences in emotion: expression, experience, and physiology. *J. Personal. Soc. Psychol.* **74**(3), 686 (1998)
52. Kveraga, K., Boshyan, J., Bar, M.: Magnocellular projections as the trigger of top-down facilitation in recognition. *J. Neurosci.* **27**(48), 13232–13240 (2007)
53. Kveraga, K., Ghuman, A.S., Bar, M.: Top-down predictions in the cognitive brain. *Brain Cogn.* **65**(2), 145–168 (2007)
54. Kveraga, K., Ghuman, A.S., Kassam, K.S., Aminoff, E.A., Hämäläinen, M.S., Chaumon, M., Bar, M.: Early onset of neural synchronization in the contextual associations network. *Proc. Natl. Acad. Sci.* **108**(8), 3389–3394 (2011)
55. Kveraga, K., Boshyan, J., Adams Jr, R.B., Mote, J., Betz, N., Ward, N., Hadjikhani, N., Bar, M., Barrett, L.F.: If it bleeds, it leads: separating threat from mere negativity. *Soc. Cogn. Affect. Neurosci.* **10**(1), 28–35 (2015)
56. Kveraga, K., Im, H.Y., Ward, N., Adams Jr, R.B.: Fast saccadic and manual responses to faces presented to the koniocellular visual pathway. *J. Vis.* **20**(2), 9–9 (2020)
57. Langton, S.R.: Gaze perception and visually mediated attention. In: Adams, R., Ambady, N., Nakayama, K., Shimojo (eds.) *The Science of Social Vision*, pp. 108–132. Oxford University Press, Oxford (2010)
58. Larsen, R.J., Diener, E.: Promises and problems with the circumplex model of emotion. In: Clark, M.S. (ed.) *Emotion*, pp. 25–59. Sage Publications (1992)
59. Lewis, G.J., Lefevre, C.E., Bates, T.C.: Facial width-to-height ratio predicts achievement drive in us presidents. *Personal. Individ. Differences* **52**(7), 855–857 (2012)
60. MacDonell, E.T., Geniole, S.N., McCormick, C.M.: Force versus fury: Sex differences in the relationships among physical and psychological threat potential, the facial width-to-height ratio, and judgements of aggressiveness. *Aggressive Behav.* **44**(5), 512–523 (2018)
61. Mansour, C., Haier, R., Buchsbaum, M.: Gender comparisons of cerebral glucose metabolic rate in healthy adults during a cognitive task. *Personal. Individ. Differences* **20**(2), 183–191 (1996)
62. McManus, C.: Half a century of handedness research: Myths, truths; fictions, facts; backwards, but mostly forwards. *Brain Neurosci. Adv.* **3**, 2398212818820513 (2019)
63. Mishkin, M., Ungerleider, L.G.: Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behav. Brain Res.* **6**(1), 57–77 (1982)
64. O'Callaghan, C., Kveraga, K., Shine, J.M., Adams Jr, R.B., Bar, M.: Convergent evidence for top-down effects from the “predictive brain”. *Behav. Brain Sci.* **39**, e254 (2016)
65. Öhman, A., Flykt, A., Esteves, F.: Emotion drives attention: detecting the snake in the grass. *J. Exp. Psychol. Gen.* **130**(3), 466 (2001)
66. Orozco, S., Ehlers, C.L.: Gender differences in electrophysiological responses to facial stimuli. *Biol. Psychiatry* **44**(4), 281–289 (1998)

67. Richardson, T., Gilman, R.T.: Left-handedness is associated with greater fighting success in humans. *Sci. Rep.* **9**(1), 15402 (2019)
68. Rogers, L.J., Vallortigara, G., Andrew, R.J.: *Divided Brains: The Biology and Behaviour of Brain Asymmetries*. Cambridge University Press (2013)
69. Russell, J.A.: A circumplex model of affect. *J. Personal. Soc. Psychol.* **39**(6), 1161 (1980)
70. Russell, R.: Why cosmetics work. In: *The Science of Social Vision*, vol. 7, pp. 186–204. Oxford University Press, Oxford, England (2011)
71. Sacco, D.F., Brown, M., Lustgraaf, C.J., Young, S.G.: Women’s dangerous world beliefs predict more accurate discrimination of affiliative facial cues. *Evol. Behav. Sci.* **11**(4), 309 (2017)
72. Shaywitz, B.A., Shaywitz, S.E., Pugh, K.R., Constable, R.T., Skudlarski, P., Fulbright, R.K., Bronen, R.A., Fletcher, J.M., Shankweiler, D.P., Katz, L., et al.: Sex differences in the functional organization of the brain for language. *Nature* **373**(6515), 607–609 (1995)
73. Stevens, J.S., Hamann, S.: Sex differences in brain activation to emotional stimuli: a meta-analysis of neuroimaging studies. *Neuropsychologia* **50**(7), 1578–1593 (2012)
74. Stokes, M.B., Payne, B.K.: Mental control and visual illusions: Errors of action and construal in race-based weapon misidentification. In: *The Science of Social Vision*, pp. 295–305. Oxford University Press (2010)
75. Thayer, R.E.: Moods of energy and tension that motivate. In: *The Oxford Handbook of Human Motivation*, p. 408. Oxford University Press (2012)
76. Tsujimura, H., Banissy, M.J.: Human face structure correlates with professional baseball performance: insights from professional Japanese baseball players. *Biol. Lett.* **9**(3), 20130140 (2013)
77. Van Le, Q., Isbell, L.A., Matsumoto, J., Nguyen, M., Hori, E., Maior, R.S., Tomaz, C., Tran, A.H., Ono, T., Nishijo, H.: Pulvinar neurons reveal neurobiological evidence of past selection for rapid detection of snakes. *Proc. Natl. Acad. Sci.* **110**(47), 19000–19005 (2013)
78. Ward, N., De Vito, D., Cushing, C., Boshyan, J., Im, H.Y., Adams Jr, R., Kveraga, K.: Neurodynamics and hemispheric lateralization in threat and ambiguous negative scene recognition. *J. Vis.* **17**(10), 313–313 (2017)
79. Watson, D., Tellegen, A.: Toward a consensual structure of mood. *Psychol. Bull.* **98**(2), 219 (1985)
80. Windhager, S., Schaefer, K., Fink, B.: Geometric morphometrics of male facial shape in relation to physical strength and perceived attractiveness, dominance, and masculinity. *Am. J. Hum. Biol.* **23**(6), 805–814 (2011)
81. Zebrowitz, L.A.: Ecological and social approaches to face perception. In: *Oxford Handbook of Face Perception*, pp. 31–50 (2011)
82. Zilioli, S., Sell, A.N., Stirrat, M., Jagore, J., Vickerman, W., Watson, N.V.: Face of a fighter: Bzygomatic width as a cue of formidability. *Aggressive Behav.* **41**(4), 322–330 (2015)

Chapter 6

From Pixels to Power: Critical Feminist Questions for the Ethics of Computer Vision



Flora Oswald

Abstract As computer and robotic applications are empowered with new visual capabilities, the promises and pitfalls of these technologies are increasingly salient. While many researchers and data scientists are rightly enthused by the innovative capacity of these technologies, others have called attention to the ways in which these technologies can reinforce intersecting systems of social, economic, and political oppression. In this chapter, I describe prominent concerns from the human side of computer vision, including algorithmic bias, invisible labor forces, and the socially (ir)responsible application of computer vision technology in a socially inequitable world. In doing so, I draw upon descriptions of artistic resistance, emotional labor, and feminist care in order to flip the perspective; to understand not only how computer technologies can “perceive” aesthetics, emotion, and art, but how human expressions of these features are in an ongoing conversation with computer vision technologies. I engage intersectional and standpoint feminisms to open a discussion of power, oppression, and potential futures in computer vision.

6.1 Introduction

As outlined in the pages of this volume, there are great potential promises of computer vision and algorithmic modelling for understanding and enhancing the human experience. However, many are critical of these technological developments; the literature on ethical challenges in algorithmic systems follows technological advances in lockstep. Noted concerns include embedded biases against racial [33, 37] and gender minorities [14, 18], invisible and exploitative labor practices hidden by the opaque nature and alleged neutrality of algorithmic models [14], and potential nefarious uses, such as surveillance of migrants [35] and detection of queer faces [7].

F. Oswald (✉)

Department of Psychology, University of Connecticut, Storrs, CT, USA

e-mail: flora.oswald@uconn.edu

Notably, the burden of these biases is borne by those who already contend with marginalization [15].

In this chapter, I contend with ethical issues on the human side of computer vision, extending upon human-centered [3] and data feminist approaches [14]. Furthermore, I draw upon descriptions of artistic resistance, emotional labor, and feminist care in order to flip the perspective; to understand not only how computer technologies can “perceive” aesthetics, emotion, and art, but how human expressions of these features are in an ongoing conversation with computer vision technologies. My specific orientation to ethical issues in computer vision draws upon intersectional [11, 12] and standpoint feminisms [10, 21–23] to understand who is left out of historical and contemporary computer vision [39] and how patterns of leaving out reflect existing systems of oppression. Feminist scientists have long grappled with issues of systemic bias and structural inequalities and thus feminist approaches provide a critical perspective from which to analyze the ethics of computer vision; see also [15].

It is worthy to note here my own positionality and how it shapes my approach to these issues. I write as a queer White cisgender woman, and as someone with an advanced degree. I am trained as a feminist social scientist, and my work considers power dynamics in intergroup relations, particularly in intergroup perception; the overarching goal of my work is to understand how people with marginalized identities visually perceive their social worlds; see [34]. As a social scientist who studies visual perception and social inequality, I engage with issues I describe throughout this chapter—including data management, stimulus development, and institutionalized bias—through a feminist lens in my own work. I cannot claim to have experienced all the forms of marginalization which I touch upon in this chapter. Taking an intersectional perspective allows me to discuss these inequalities, but I hope not to obscure the voices of those who directly experience them.

6.2 Feminist Approaches

6.2.1 *Intersectional Feminisms*

Intersectionality calls to attention the multidimensionality of marginalized subjects’ experiences [9, 11, 12] to reject single-axis constructions of identity and instead forward a more complex notion of identity such that “subjectivity is constituted by mutually reinforcing vectors of race, gender, class, and sexuality” [32, p. 2]. Though numerous social identities have implications for subjectivity, intersectionality has historically focused on, and is “...particularly adept at capturing and theorizing the simultaneity of race and gender as social processes” and “from its inception...has had a long-standing interest in one particular intersection: the intersection of race and gender” [32, p. 2].

Intersectional feminist approaches encourage analysis of how multiple differing vectors of oppression interact to shape hierarchies of power and those subject to them. At its core, intersectional feminism analyzes unequal systems of power, and therefore offers a lens through which to analyze how these systems of power shape, and are shaped by, emerging technologies including computer vision.

6.2.2 *Standpoint Feminisms*

Standpoint theory contends that all individuals have a limited perspective on the world which is shaped by their social location and experiences. This limited perspective comprises their knowledge of the world as it is situated in and by their social location. Individuals have embodied knowledge—indeed, embodied perceptions of the world—shaped by their experiences, including by their belonging in communities with certain locations in social hierarchies [10, 21, 23]. Each individual thus has a partial perspective shaped by their unique social positionality.

Standpoint approaches reject a “view from nowhere” [22]; a view often characteristic of the sciences as they are constructed as objective and removed from their human creators. Indeed, algorithms—including those involved in computer vision—are often similarly construed as objective and autonomous, as operating outside of human direction (see [8]). Yet, as Amoore [2] describes algorithmic models, they are “always already partial accounts”. As I discuss in the following sections, these partial accounts are shaped by their human creators and the broader hierarchical social contexts in which such technologies are developed.

These hierarchies of power are particularly important to understand when deconstructing the notion of autonomous and objective algorithmic models, particularly because these models tend to be shaped by the ideologies and contexts of those with relative social power, and are thus biased to maintain existing hierarchies of power. It is thus important to understand the contexts in which these technologies are developed, and by whom they are developed, in order to understand how they may be shaped by their social locations. Standpoint theory offers that marginalized social positions have the potential to generate perspectives that are “less partial and less distorted” than those generated by dominant social positions [22]. I draw upon the potential of marginalized social positions to offer a new and less distorted perspective, and the understanding of algorithmic models in computer vision as partial perspectives, in order to argue for both greater understanding of the role (and ethics) of human involvement in computer vision processes, and for the need for greater diversity in the humans involved.

6.3 Algorithmic Bias

Algorithmic bias—systemic errors in algorithms which create predictably unfair or imbalanced outcomes—has been a prominent concern, making headlines in popular media and giving rise to general distrust in artificial intelligence applications. Though algorithmic bias does not necessarily reflect societal biases, many forms of algorithmic bias do indeed replicate social biases such as racism and sexism. Because many algorithms, especially from the end user’s point of view, are a black box—that is, for reasons of complexity, intellectual or corporate ownership, and more, we often do not know how a given algorithm was designed or trained, or exactly how it works (not to mention the contrary notion that algorithms are objective and removed from human influence)—it can be especially hard to detect these biases.

However, many have begun to do so. High-profile examples of detected biases include computer vision algorithms noted to produce outcomes such as misclassifying images of Black people as non-human primates (e.g., chimpanzees and gorillas; [1]), computer vision systems designed to detect pedestrians failing at higher rates at detecting those with darker skin tones [42] and gender classification systems performing better for lighter-skin men than darker-skinned women [5]; such systems also often neglect those who fall outside of binaries (e.g., trans and gender diverse people; those with plurisexual identities; [7, 24]).

These biases emerge, in large part, due to the data upon which the algorithms were trained. For example, [5] reveal that only a very small percentage of training data for many facial-analysis softwares are images of dark-skinned women. This results in outcomes such as Ghanian-American then-graduate student Joy Buolamwini’s experience—which prompted the analysis of training images—in which she discovered that a facial-analysis software could not detect her dark-skinned face (unless she put on a white mask). That these limited and non-representative training data were not considered problematic by the developers of the software demonstrates what D’Ignazio and Klein [14] call a privilege hazard, whereby those with social power are ill-equipped to detect or recognize instances of oppression. That is, those who developed the software were not cognizant (or worse, were cognizant but chose to ignore) that their application did not account for and indeed would not work for those with darker skin. Their computer vision software thus took a limited perspective: one which privileged Whiteness.

Such privilege hazards are ingrained in many computer vision algorithms, both by the limited demographics of developers (typically well-educated, elite men and other dominant group members; [14]) and by the limited data on which they are trained. In effect, computer vision models prioritize a certain perspective: that trained into them by their developers. Computer vision algorithms thus have “not the capacity to ‘see’ but that of making judgements over what in the visual world should be seen and how” [35]; the “what” and “how” are informed by the partial perspectives of their developers (who, again, tend to come from relative social privilege). Though these biases emerge at the level of the individual, the collective

hazard of computer vision’s domination by elite White men means that these biases reach “the hegemonic, disciplinary, and structural domains as well” [14].

As I have argued elsewhere [34], prioritizing one group of perceivers—especially a socially privileged group—severely limits the generalizability of our knowledge, and, by extension, that of our computer vision models. Who makes the decision about what computer vision algorithms should see, and how they should see it, has potentially devastating consequences such as those reviewed above. Given who is making these decisions and their own partial perspectives, these consequences tend to be borne by those who are already marginalized. Too often, embedded biases are obscured by the notion of computer vision as a post-human technology (e.g., [8, 35]) and by the marginalization of those who suffer their consequences; their lack of structural and institutionalized power often means they are left out of discussions of potential harms (though, as we will see later in this chapter, they are not without creative strategies for activism and resistance).

What algorithms should see, and how they should see it, are not shaped only by their developers. There are hierarchical and very human power structures underlying how computer vision technologies are trained to “see,” and these hierarchies often themselves remain unseen.

6.4 Invisible Labor

Two forms of invisible labor are particularly relevant to computer vision. First, there is the labor of those who provide the data used to train computer vision applications; second, there are those who are employed in the task of annotating this data. When computer vision models are trained, they are typically exposed to a set of images which are already labelled—for example, they might be trained on a database of facial images, which contains information about the age, race, gender, etc. of each face. In theory, once trained on such images, the computer vision application will then be able to generalize this knowledge to detect these features in new data—thereby gaining the ability to “see” these features.

Much of the data used to train these models, however, comes with ethical challenges. For example, images are often extracted without consent or even awareness of the subjects, and without ethical approval [4, 40]. People are thus unwillingly represented in large-scale training datasets, sometimes for projects with potential harms to their own marginalized group (therefore, not likely somewhere they would like to be represented). Ref. [7] critiques this kind of data harvesting with regard to development of a computer vision gaydar system [41], which they purported to be more accurate than human judges at detecting sexual orientation from facial images. In this study, the face stimuli were harvested images of users of Facebook and dating apps; it is not clear whether such images are technically public data and available for research use, but moral and ethical concerns about privacy abounded in the wake of the publication (e.g., [7, 13]).

Furthermore, concerns were noted about how these images were being used: In an algorithm that could potentially be used to identify and trace gay and lesbian people (the system could only distinguish between heterosexual and gay/lesbian orientation in a binary fashion) with intent to harm (see [7, 13]). Gay and lesbian subjects were thus unwittingly involved in the development of a computer vision application with the potential for significant harm to their own group, and without the potential for significant benefit [13]. Though the training data included only White, U.S. faces, the authors reason that their findings should extend to other groups, thus the potential harm extends beyond even those included in the actual data. Others have detailed how face harvesting has also been implemented on the facial data of immigrants, abused children, and dead people [25]. Though this harvesting of one’s data for potentially nefarious means may not represent “labor” as we typically conceive of it, I argue that these individuals are being exploited at great potential cost to themselves and materials they have produced are being utilized without their consent (while, for example, others are paid for their images or participation in creating facial databases), which can also incur emotional work to members of marginalized groups who recognize, resist, or must manage the resultant outcomes of these processes.

The development of computer vision technology not only relies on images of often-unwilling participants, but also on exploitative human labor practices of labelling and annotating images. In order for computer vision to detect features of images, algorithms must be trained on datasets of images in which these features are already labelled; this labelling is crowdsourced from underpaid and undervalued workers, often in developing countries (e.g., [14, 28]). This labor is often described as Ghost Work [19], indicating its invisibility in the ostensibly objective and autonomous development of computer vision. That is, this labor is obscured by the predominant conceptualization of artificial intelligence as post-human, which necessarily conceals the role—and subjectivity—of human labor in developing these technologies.

That training data are curated by humans results in data that “necessarily reflect parts of their knowledge, assumptions, and values and their socio-cultural-political context in general” [15, p. 329]. Some have argued that the subjectivities—that is, the partial perspectives—of the workers involved in labelling and annotating data interfere with the potential objectivity of computer vision models (see [28] for review). However, annotator subjectivities are often subdued by top-down impositions of the task (i.e., instructions from the client requesting the work) and annotators may not even be aware of the sources of the images and purpose of project their labor contributes to [28]. It is thus important to consider, at the structural and institutional level, how systems of oppression result in those with power having the right to impose meaning on the data through the instructions they pass down the hierarchical labor chain.

Ref. [29] describe data as resulting from these systems of oppression that “are present among data workers as well as in the relationship between those whose data is collected and those who make use of data for research and/or profit” (p. 161), highlighting the multiple intersecting vectors of capitalist power involved in

producing data. In an analysis of computer vision dataset annotation, [28] reveal how workers engaged in data annotation look to those with greater structural power to shape their annotation decisions, revealing how the perspectives of companies and clients on what should be seen in data trickle down through the ranks, resulting in a hegemonic form of “vision” that prioritizes the perspectives of those with power. Thus, barring even application, computer vision models are embedded into oppressive power structures early in their development. Training data is often privacy-compromised and fails to acknowledge the positionalities of who it represents (a precursor to irresponsible and potentially harmful application), and is shaped by hegemonic biases as a result of the power hierarchies surrounding the labor of data annotation. These forms of invisible labor have given rise to another form of labor, one with the potential to create a more socially just future for computer vision and artificial intelligence more broadly: activist labor.

6.5 Resistance

Activism is a way to antagonise the current hegemonic structures involved in computer vision and to “actively envision how we can center other values and paradigms” [35, p. 34]. Many forms of activist resistance are rooted in marginalized peoples’ rejections of these oppressive systems; “most marginalized users are not waiting for Big Tech to deliver the solution... communities may not have the time to wait while they are actively experiencing harms” [17, p. 98]. Activist movements around artificial intelligence and machine learning, as well as computer vision in particular, have employed a diverse range of strategies to resist and re-envision algorithmic justice.

In this section, I will first describe some recent technical advances within the traditional boundaries of AI/computer vision which have potential to counter ethical quandaries, and will then describe resistance by activist movements and their approaches. In particular, I focus on artistic, emotional, and aesthetic expressions of resistance, in line with the overarching themes of this book. Specifically, I aim to illustrate how these human expressions can illuminate the human side of artificial intelligence technologies, provoking a more just and equitable sociotechnical future.

6.5.1 *Responses Within*

Ethics is an active field within artificial intelligence, and the artificial intelligence community has sought to develop technical solutions to its own ethical conundrums. For example, artificial intelligence ethics initiatives have resulted in a proliferation of principles documents and guidelines for the ethical use of AI (see [6, 20, 31]); such guidelines offer conceptual-level principles for engaging ethically with AI technologies. However, critiques of such guidelines indicate that they are too

abstract and shy away from methodological recommendations, resulting in a proliferation of conceptual documents without a corresponding rise in practical implementation (e.g., [20, 27]).

Despite this, recent practical advances in computer vision technology demonstrate potential to address ethical problems. For example, [16] outline recent advances in bias discovery and quantification as well as developments in the collection of bias-aware datasets. Although the authors identify remaining space for improvement in bias detection and reduction, their review highlights a number of recent technological developments with the potential to reduce bias in novel algorithms and detect bias in existing algorithms. Another recent development with the potential to reduce power inequities related both to invisible labor and to stimulus harvesting and privacy is the development of improved strategies for building adaptive pre-trained models from advanced synthetic data (e.g., [30]). Models based on synthetic training data sidestep the need for images of real people, and thus the associated ethical concerns. While synthetic data has long been considered as an avenue for addressing these issues, only with recent advances in synthetic data development can such synthetic data be optimally leveraged to develop transferable models [30].

While these technological advancements represent changes in the capacity of computer vision technologies to be developed and applied in ethical ways, activist responses offer a different, provocative lens through which to understand resistance to computer vision technologies and their ethical constraints, as well as one through which different computer vision futures can be imagined.

6.5.2 Activist Responses

One predominant example of activist resistance is the Algorithmic Justice League, founded by Joy Buolamwini (mentioned earlier, who discovered that facial analysis technology could not detect her face as a result of their biased training on White images). The Algorithmic Justice League “is an organization that combines art and research to illuminate the social implications and harms of artificial intelligence” [26]; for example, through the production of visual poetry, TED talks, and a recent film highlighting racial and gender bias in artificial intelligence. These expressions engage the public in the practice of caring about and for these issues, and allows the public to contribute records of their own experiences with biased artificial intelligence. In this way, the Algorithmic Justice League embraces a feminist ethic of care—listening to others in the public, responding to their needs through activism, and focusing on and valuing the lives and experiences of other humans as they intersect with imbalances of power (see [38]). The Algorithmic Justice league offers a collaborative, community-driven approach to highlight the voices of those marginalized and harmed by biased artificial intelligence, with the intention of driving policy change directed toward a more equitable future.

Others have developed creative technologies to counter the ethical issues of data harvesting. For example, the artistic virtual environment PeopleSansPeople offers a mode of training human-centric computer vision models without human data. PeopleSansPeople is a synthetic data generator that creates 3D scenes populated by relatively diverse 3D models with several parameters for variation. The ability to manipulate the environment, people, and poses in the simulator bypasses the privacy and ethical concerns associated with human data, and allows for heightened diversity in datasets. Though this technology could still be used for nefarious means, it eliminates one set of concerns with data sourcing and thus offers the potential for a more equitable future.

More traditional forms of artistic expression have also been harnessed to reveal the human side of computer vision. For example, Kate Crawford and Trevor Paglen's *Training Humans* photography exhibition displays images and accompanying labels from computer vision training datasets to tell the story of how humans are harvested and classified in artificial intelligence systems. The goals of the project were to "engage with the materiality of artificial intelligence, and to take those everyday images seriously as a part of a rapidly evolving machinic visual culture. That required us to open up the black boxes and look at how these 'engines of seeing' currently operate" [36]. The artists were particularly interested in exploring how human emotions and facial expressions are classified in training datasets, and in how these classifications relate to facial analysis software's assessments of mental health, hireability, criminality, and other morally laden judgements. The project raised awareness of the ethical issues and biases embedded in these training systems, eventually resulting in ImageNet, one of the primary training image databases, removing over half a million images. This work reveals the power of artistic activism to enact change and raise awareness of power and bias in computer vision technology.

Together, these examples—and many more endeavours I do not discuss here—demonstrate the potential of human-centered artistic expression to illuminate bias and power hierarchies in computer vision. These activist efforts offer consciousness-raising opportunities and spaces to develop an ethic of care to resist the hegemonic state of computer vision and discover creative alternative futures for computer vision by re-introducing humanness to the system.

6.6 Conclusion

In this chapter, I have outlined how capitalism and sexism and racism and classism and geographical bias, among others, intersect to shape computer vision realities. I have described how bias is embedded into algorithms as a result of the partial perspectives of those who develop them, and how this results in further harm to marginalized groups. Furthermore, I have outlined the power dynamics involved in the invisible labor underlying computer vision technology, with marginalized people contributing—often in exploitative or non-consensual ways—to technologies with

great capacity to further marginalize. Throughout the chapter, I have also pointed to how the notion of computer vision as objective, autonomous, and post-human can obscure these hierarchies of power. Yet, there is hope for socially just computer vision. There are potentially responsible applications of these technologies, and many marginalized people and communities have undertaken efforts to improve computer vision technologies and to intervene, through resistance and refusal, in the harms of these technologies. I described artistic, aesthetic, and emotional forms of activism and development that have been undertaken as resistance to dominant ways of doing computer vision. The humanness of these approaches offers a radical paradigm shift to computer vision, which generates new ways of doing and caring for the human side of these technologies. Potential ethical futures of computer vision, and of artificial intelligence more broadly, must pay heed to their human side.

Feminist Questions to Consider Moving Forward

To contribute to this effort, I offer in closing a number of critical, feminist questions to ask when engaging algorithms, both in computer vision and beyond. Many of these questions are adapted from [15] and [35] who offer these (and more) in comprehensive reviews of algorithmic bias and refusal.

Which perspectives are being valued?

What power hierarchies are in place, and could inequalities be reduced?

What values are reflected in the data, and in what the model is intended to “see”?

Who will be affected by new systems of technology? do they consent to these effects?

Who is participating in the design of new systems?

Who do systems benefit? who could they harm?

What alternative paradigms or ways of approaching the problem/solution could be considered?

What forms of resistance are available? who are they available to?

References

1. Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J.W., Brundage, M.: Evaluating clip: towards characterization of broader capabilities and downstream implications. arXiv preprint arXiv:2108.02818 (2021)
2. Amoore, L.: Cloud Ethics: Algorithms and the Attributes of Ourselves and Others. Duke University Press, Durham (2020)
3. Aragon, C., Guha, S., Kogan, M., Muller, M., Neff, G.: Human-Centered Data Science: An Introduction. MIT Press, Cambridge (2022)
4. Birhane, A., Prabhu, V.U.: Large image datasets: a pyrrhic win for computer vision? In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1536–1546. IEEE (2021)
5. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91. PMLR (2018)

6. Chi, N., Lurie, E., Mulligan, D.K.: Reconfiguring diversity and inclusion for ai ethics. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 447–457 (2021)
7. Chun, W.H.K.: *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. MIT Press, Cambridge (2021)
8. Ciston, S.: Imagining intersectional AI. In: Conference on Computation, Communication, Aesthetics, & X (2019). <http://2019.xcoax.org/pdf/xCoAx2019-Ciston.pdf>
9. Collective, C.R.: The Combahee river collective statement. In: Home Girls: A Black Feminist Anthology, vol. 1, pp. 264–274 (1983)
10. Collins, P.H.: *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, New York (2022)
11. Crenshaw, K.: Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, p. 139 (1989)
12. Crenshaw, K.: Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.* **43**, 1241 (1990)
13. De Block, A., Conix, S.: Responsible dissemination in sexual orientation research: the case of the AI ‘gaydar’. *Philos. Sci.* **89**, 1–18 (2022)
14. D’ignazio, C., Klein, L.F.: *Data Feminism*. MIT Press, Cambridge (2020)
15. Draude, C., Klumbyte, G., Lücking, P., Treusch, P.: Situated algorithms: a sociotechnical systemic approach to bias. *Online Inf. Rev.* **44**(2), 325–342 (2020)
16. Fabbrizzi, S., Papadopoulos, S., Ntoutsis, E., Kompatsiaris, I.: A survey on bias in visual datasets. *Comput. Vis. Image Underst.* **223**, 103552 (2022)
17. Ganesh, M.I., Moss, E.: Resistance and refusal to algorithmic harms: varieties of ‘knowledge projects’. *Media Int. Aust.* **183**(1), 90–106 (2022)
18. Gehl, R.W., Moyer-Horner, L., Yeo, S.K.: Training computers to see internet pornography: gender and sexual discrimination in computer vision science. *Telev. New Media* **18**(6), 529–547 (2017)
19. Gray, M.L., Suri, S.: *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books, Boston (2019)
20. Hagendorff, T.: A virtue-based framework to support putting AI ethics into practice. *Philos. Technol.* **35**(3), 55 (2022)
21. Haraway, D.: Situated knowledges: the science question in feminism and the privilege of partial perspective. *Fem. Stud.* **14**(3), 575–599 (1988)
22. Harding, S.: *Whose Science? Whose Knowledge?: Thinking from Women’s Lives*. Cornell University Press, Ithaca (1991)
23. Hartsock, N.C.: *Money, Sex, and Power: An Essay on Domination and Community*. Longman, New York (1983)
24. Keyes, O.: The misgendering machines: Trans/HCI implications of automatic gender recognition. In: Proceedings of the ACM on Human-Computer Interaction (CSCW), vol. 2, pp. 1–22 (2018)
25. Keyes, O., Stevens, N., Wernimont, J.: The government is using the most vulnerable people to test facial recognition software. *Slate Mag.* **17** (2019)
26. League, A.J.: Mission, team, and story. <https://www.ajl.org/about> (n. d.)
27. McLennan, S., Lee, M.M., Fiske, A., Celi, L.A.: AI ethics is not a panacea. *Am. J. Bioethics* **20**(11), 20–22 (2020)
28. Miceli, M., Schuessler, M., Yang, T.: Between subjectivity and imposition: power dynamics in data annotation for computer vision. In: Proceedings of the ACM on Human-Computer Interaction (CSCW2), vol. 4, pp. 1–25 (2020)
29. Miceli, M., Yang, T., Naudts, L., Schuessler, M., Serbanescu, D., Hanna, A.: Documenting computer vision datasets: an invitation to reflexive data practices. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 161–172 (2021)
30. Mishra, S., Panda, R., Phoo, C.P., Chen, C.F.R., Karlinsky, L., Saenko, K., Saligrama, V., Feris, R.S.: Task2sim: towards effective pre-training and transfer from synthetic data. In: Proceedings

- of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9194–9204 (2022)
- 31. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci. Eng. Ethics* **26**(4), 2141–2168 (2020)
 - 32. Nash, J.C.: Re-thinking intersectionality. *Feminist Rev.* **89**(1), 1–15 (2008)
 - 33. Noble, S.U.: Algorithms of oppression. In: *Algorithms of Oppression*. New York University Press, New York (2018)
 - 34. Oswald, F., Adams Jr., R.B.: Feminist social vision: seeing through the lens of marginalized perceivers. *Pers. Soc. Psychol. Rev.* **27**, 10888683221126582 (2022)
 - 35. Pereira, G.: Towards refusing as a critical technical practice: struggling with hegemonic computer vision. *Peer-Rev. J. About* **10**(1), 30–43 (2021)
 - 36. Prada, F.: Kate Crawford and Trevor Paglen: training humans. <https://www.fondazioneprada.org/project/training-humans/?lang=en> (2019)
 - 37. Raji, I.D.: Handle with care: lessons for data science from black female scholars. *Patterns* **1**(8), 100150 (2020)
 - 38. Ramdas, K.: Feminist care ethics, becoming area. *Environ. Plann. D: Soc. Space* **34**(5), 843–849 (2016)
 - 39. Stewart, A.J.: Doing personality research: how can feminist theories help? In: Clinchy, B.M., Norem, J.K. (eds.) *Gender and Psychology Reader*, pp. 54–68. New York University Press, New York (1998)
 - 40. Thylstrup, N.B.: Data out of place: toxic traces and the politics of recycling. *Big Data Soc.* **6**(2), 2053951719875479 (2019)
 - 41. Wang, Y., Kosinski, M.: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *J. Pers. Soc. Psychol.* **114**(2), 246 (2018)
 - 42. Wilson, B., Hoffman, J., Morgenstern, J.: Predictive inequity in object detection. arXiv preprint arXiv:1902.11097 (2019)

Part III

Computer Social Vision

Following the previous part of the book focused on human social visual perception, this next part gets more directly into methods for applying computer vision to the processing and detecting of important social and emotional information from faces, bodies, and scenes.

The first chapter, “High-Speed Joint Learning of Action Units and Facial Expressions,” details a reliable and fast method for training computers to detect discrete emotional expressions (e.g., anger, fear, sadness, joy) from the face, focusing on specific facial muscle patterns (combinations of different action units) as delineated by the well-known Facial Action Coding System (FACS).

The second chapter, “ExpressionFlow: A Microexpression Descriptor for Efficient Recognition,” presents a novel method for reading micro-expressions in the face. These are micro-momentary facial muscle patterns, as detailed in the preceding chapter, that are thought to convey genuine emotional states, and may appear while an expressor is attempting to express something else. This approach searches for geometric motion patterns from microexpression image sequences.

The third chapter, “Emotion in the Neutral Face: Applications for Computer Vision and Aesthetics,” reviews work showing how even subtle resemblance of discrete emotions (e.g., anger, fear, sadness, joy) in otherwise neutral facial displays can be algorithmically detected and applied to predict trait impressions from the face. This chapter questions whether faces can ever really be construed as “neutral” and discusses implications for computer vision and aesthetics.

The fourth chapter, “Multi-Stream Temporal Networks for Emotion Recognition in Children and in the Wild,” extends this work further by examining emotion perception from faces, bodies, and scenes. This chapter also extends the focus in this research that is typically on adult perceivers to children as well.

Chapter 7

High-Speed Joint Learning of Action Units and Facial Expressions



Feng Xu, Yifan Yuan, Junping Zhang, and James Z. Wang

Abstract Facial expressions serve as a crucial facet of human behavior, offering a wealth of social and emotional cues. Despite their significance, achieving real-time, accurate, and interpretable recognition of facial expressions from multimedia content has posed a considerable challenge for computer systems. In an effort to address these concerns, we present a novel sparse tagging-like methodology to jointly learn Action Units (AUs) and facial expressions. Our approach regards the process of AU combination recognition as image tagging, thereby significantly reducing computational complexity through the exclusive use of matrix multiplications. To enhance the interpretability of our analysis, we incorporate a sparse term into the methodology, promoting the sparseness of AU combinations. An evaluation of our proposed technique across five benchmark datasets reveals its superiority in terms of speed, interpretability, and robustness compared to existing algorithms, while maintaining commensurate levels of accuracy. This refined approach represents a significant advancement in the field of facial expression recognition and analysis, offering a more efficient solution for real-time applications.

F. Xu

School of Computer Science, Fudan University, Shanghai, China
Ant Financial Group, Hangzhou, China
e-mail: feng_xu@fudan.edu.cn

Y. Yuan · J. Zhang (✉)

School of Computer Science, Fudan University, Shanghai, China
e-mail: yfyan21@m.fudan.edu.cn; jpzhang@fudan.edu.cn

J. Z. Wang

Data Science and Artificial Intelligence Area, and Human-Computer Interaction Area, College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA
e-mail: jwang@ist.psu.edu

7.1 Introduction

As important visible manifestations of human emotional states, facial expressions often play a pivotal role in interpersonal communication [4]. Automatic recognition of facial expressions has far-reaching applications in various domains, including human-computer interaction [27], driver drowsiness detection [10], and lie detection for security and law enforcement [21]. Furthermore, computer-assisted expression recognition can enhance the social experiences of individuals with certain medical conditions that impede their ability to recognize facial expressions or emotions [22].

In the multimedia database domain, annotating multimedia content with facial expressions enables users to search for content featuring individuals in specific emotional states [31]. This functionality can be employed to query personal photo collections for images depicting surprise, for example, or to analyze interpersonal emotions in group meetings [14].

Effective recognition systems often incorporate techniques and knowledge from multimedia, computer vision, and psychology fields. Presently, expression recognition methodologies can be broadly classified into two categories: *judgment-based* and *sign-based* approaches [7]. From expression images, the former attempts to model the corresponding emotional states [12, 15], whereas the latter characterizes input expressions according to combinations of predefined signs, each reflecting a local facial action. A seminal sign-based approach is the Facial Action Coding System (FACS), developed by Ekman [6]. In FACS, the signs are referred to as Action Units (AUs). A key advantage of sign-based approaches is their enhanced interpretability when modeling facial action with AUs.

Under the sign-based paradigm, existing methods typically adopt a similar framework: feature representations are extracted either from local facial features or entire expression images, followed by the construction of classifiers for specific AUs of interest. However, these methods often overlook potential relationships among AUs. For instance, certain AUs such as *Lip Presser* (AU24) and *Lips Apart* (AU25) are mutually exclusive, yet they can be simultaneously incorrectly detected by different classifiers from the same image. Addressing such errors requires *joint learning* of AUs.

In this chapter, we introduce a novel perspective for AU recognition to address these challenges. Instead of learning separate classifiers for each AU on pre-extracted visual features, we jointly learn filters on the original expression images for tagging all AUs. Subsequently, the responses of these filters are directly used for predicting the presence of each AU. Finally, we train a single classifier to estimate the emotional state of the facial expression image based on the predicted AU combination.

The main **contributions** of our work can be encapsulated as follows:

- *Speed*: Without explicit features extraction, our method achieves magnitudes faster processing speed by employing a much lower complexity predictor.
- *Robustness*: In contrast to conventional facial expression recognition systems, our method treats AUs as tags and introduces a sparse tagging algorithm for

expression recognition, with experimental results substantiating the effectiveness of our proposed method.

- *Interpretability:* Serving as an implicit feature extractor, the filters in our method offer a more meaningful representation of AUs than manually defined visual features. The integration of a sparsity constraint further enhances the interpretability of these filters.

The remainder of this chapter is structured as follows: Sect. 7.2 provides a review of existing literature on facial expression recognition. In Sect. 7.3, we delve into the specifics of our proposed approach. Section 7.4 presents a thorough evaluation of our algorithm’s performance across five benchmark datasets. We conclude and discuss future work in Sect. 7.5.

7.2 Related Work

An important step in expression-related emotion analysis is the recognition of AUs. Given an expression image, the process aims to retrieve all AUs associated with the expression. In the sign-based approach, the task is to identify the corresponding classes or AUs associated with a given facial expression from a single image or a sequence of images. Conventional methods generally employ multiple classifiers, each designed to predict the presence of a specific AU.

Jeni et al. [13] utilized a constrained local model [23] for landmark location. For a specific AU, different image patches around landmarks are selected based on the location of different AUs. Non-negative Matrix Factorization (NMF) [16] was employed to learn filters for feature representation. A linear support vector regression (SVR) was learned to estimate AU intensity. However, this method requires the manual selection of image patches to better describe the specific AU. Furthermore, the selection of domain knowledge relies on human expertise.

Without the need for domain knowledge, the system by Littlewort et al. [18] extracted Gabor wavelet features from facial expression images, followed by Adaboost [9] feature selection [1]. Each AU was recognized separately using a support vector machine (SVM). The distance between the expression sample and the decision boundary indicates the strength of the corresponding AU. Although Gabor filters attain better performance, they generate very high dimensional features that required dimensionality reduction for subsequent recognition. Meanwhile, nonlinear SVM classifiers incurred high computational costs and low interpretability.

Compared with high-dimensional Gabor features, low-dimensional descriptors are favored by researchers in the facial expression recognition area because they offer more intuitive explanations without the involvement of dimensionality reduction. For instance, Chew et al. [3] utilized both Local Binary Pattern (LBP) and pixel-based representations for AU classification. A constrained local model is used to locate predefined landmarks on input expression images. Based on these landmarks, facial images are aligned to reduce the influence of a person’s identity.

A set of separate SVM classifiers are also employed to predict the AU combination. However, both [3] and [13] rely on accurate landmark location, which remains an open problem.

In addition to the previously mentioned static image-based methods, movement information present in video data has also been exploited. Upon segmenting input faces into several distinct regions, the method by Shreve et al. [26] estimated optical flow within each region for calculating an optical strain, which serves as an indicator of movement intensity. In general, higher values of optical strain imply the activation of a specific AU. However, this approach is sensitive to face alignment and the division scheme, requiring specialized domain knowledge. Zhang et al. [34] implemented local Gabor features in consecutive frames to facilitate movement matching, thereby establishing a discriminative representation for facial expression recognition. Nonetheless, these methods incorporate movement matching, which is both inefficient and unsuitable for real-time applications.

Automatic feature learning has been utilized as well. For instance, Long et al. [19] applied Independent Component Analysis (ICA) to learn a filter set from other datasets for the purpose of extracting facial features. Wang et al. [32] utilized Discriminant Tensor Subspace Analysis (DTSA) to learn projections that optimize both within- and between-class distances. Despite the fully automatic nature of this method and its lack of reliance on specific knowledge, it exhibits a slow convergence rate and requires substantial storage capacity.

In light of the advantages and drawbacks of these existing techniques, it is imperative to develop a facial expression recognition system that possesses strong interpretability and rapid computational speed, in addition to competitive accuracy. In the present study, we seek to utilize image tagging techniques to achieve this objective. Broadly defined, image tagging involves predicting *tags* for a given image, which may be used to describe the theme of the picture or objects in it [5, 8, 17]. For instance, FastTag [2] introduced a fast image tagging algorithm that predicts tags through straightforward matrix multiplication. To address the issue of incomplete tags, the algorithm assumes a loss of tags in the training set and compensates for it with enrichment.

Translating image tagging research into AU recognition, a simple yet intuitive perspective is to view the prediction of AUs as the assignment of tags to a given expression image. As only a limited number of AUs exist in a specific expression image, the AUs associated with a particular expression are sparse and incomplete, similar to the image tagging process. Such a relationship inspired us to investigate facial expression recognition based on image tagging, an area that has yet to be explored within the existing literature.

7.3 The Algorithm

In this section, we describe our approach for AU prediction and emotion recognition in detail. For better understanding, the whole workflow is illustrated in Fig. 7.1. From the figure, it can be seen that each filter M_i learned from the trained set

corresponds to a unique AU. When these filters are separately applied to input expression, the responses indicate the presence of corresponding AU. AU prediction results are then used as input data for the next layer for inferring the emotional state. Note that the emotion inference procedure is not directly executed on the raw AU prediction, which will be explained in detail in Sect. 7.3.3. Sections 7.3.1 and 7.3.2 are devoted to the problem formulation and optimization, respectively.

7.3.1 Problem Formulation

Assuming that the i -th face image of size $m \times n$ in the training data has been cropped and aligned, we can vectorize it to $X_i \in \mathbb{R}^{mn \times 1}$, denoting the i -th image of the training data. The training data is used to learn a filter M_t such that the response $X_i^T M_t$ directly predicts the presence of AU_t in X_i . A more effective way is to stack the filters M_t into a matrix \mathbf{M} so that $X_i^T \mathbf{M}$ can directly predict the AU combination. This naturally leads to the optimization function

$$\mathbf{M} = \arg \min_{\mathbf{M}} \frac{1}{N} \sum_{i=1}^N \|X_i^T \mathbf{M} - U_i\|^2, \quad (7.1)$$

where $U_i \in \{0, 1\}^L$ is the corresponding AU combination of X_i . L is the number of AUs of interest and N is the number of images. Each entry $U_{it} \in \{0, 1\}$ denotes the presence of the t -th AU (i.e. AU_t) of the i -th input image. This workflow is illustrated in the left part of Fig. 7.1.

It is noted that practical expressions generally involve a few AUs, which means that most elements in U_i are zeros. Equivalently, the vector of \mathbf{U} is sparse. Such a

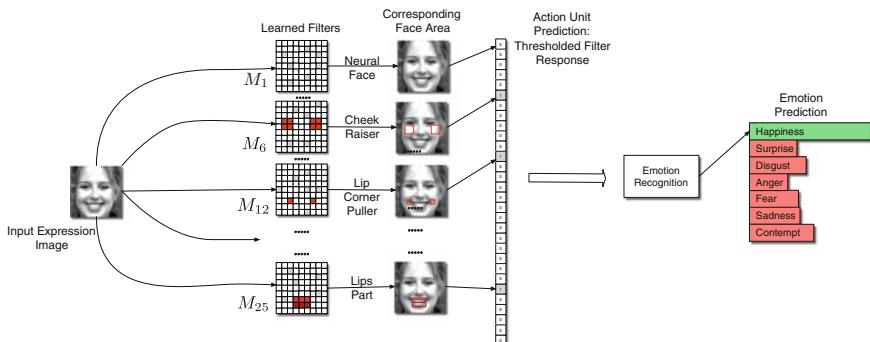


Fig. 7.1 Workflow of the algorithm. M_i denotes a learned filter associated with the i -th AU. The filters are applied to input expression, and responses indicate the presence of corresponding AU. Then a linear classifier is used to predict the emotional state based on the activation of AU prediction. Detail of emotion inference will be discussed in Sect. 7.3.3

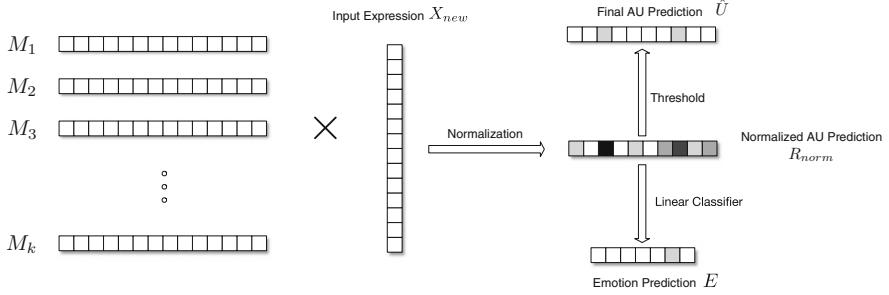


Fig. 7.2 The model training process. The major part we try to minimize is the difference between AU prediction $X^T \mathbf{M}$ and enriched AU combination $U^T \mathbf{B}$. In addition to standard ℓ_2 regularizer, we assume a sparsity relationship between expression images and their AUs, which leads to the $\|\mathbf{M}\|_1, r(\mathbf{B})$ regularize \mathbf{B} to best reconstruct AUs

sparsity phenomenon leads to the number of negative AUs ($U_{it} = 0$) dominating the number of AUs. As a result, some positive AUs can be misclassified to negative ones in high probability due to this imbalance distribution. This may deteriorate the subsequent expression recognition.

To address this issue, inspired by [2], a compensator matrix \mathbf{B} is introduced to enrich the U_i . Specifically, we aim at finding the stacked filters \mathbf{M} as well as the enrich operator \mathbf{B} , such that $X_i^T \mathbf{M}$ predicts the enriched AU combination $U_i^T \mathbf{B}$. That is, it is necessary to minimize $\|X_i^T \mathbf{M} - U_i^T \mathbf{B}\|$ during the training stage, which is shown in the center in Fig. 7.2.

Because of the imbalanced distribution, however, it is difficult to exactly capture the presumed complete AUs. One trade-off is to perform the corruption on current AU combinations, and to reconstruct them using a compensator matrix \mathbf{B} . Such a matrix \mathbf{B} , when applied to current AU combinations, would approximate the original complete ones. Formally, the following reconstruction error should be minimized with respect to \mathbf{B} :

$$\mathbf{B} = \arg \min_{\mathbf{B}} \frac{1}{N} \sum_{i=1}^N \|U_i - \tilde{U}_i^T \mathbf{B}\|^2, \quad (7.2)$$

where \tilde{U}_i is a corrupted version of U_i , with a probability p , such that $P(\tilde{U}_{it} = 0) = p$ and $P(\tilde{U}_{it} = U_{it}) = 1 - p$. The expected reconstruction error under corruption distribution is given by

$$r(\mathbf{B}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left[\|U_i - \tilde{U}_i^T \mathbf{B}\|^2 \right]_{p(\tilde{U}_i|U_i)}, \quad (7.3)$$

where $\mathbb{E}(\cdot)$ is the mathematical expectation. We can rewrite the expected loss in a closed form

$$r(\mathbf{B}) = \frac{1}{N} \text{trace}(\mathbf{B}\mathbf{T}\mathbf{B}^T - 2\mathbf{S}\mathbf{B}^T + \mathbf{U}\mathbf{U}^T), \quad (7.4)$$

where

$$\begin{aligned} \mathbf{S} &= (1-p)\mathbf{U}\mathbf{U}^T, \\ \mathbf{T} &= (1-p)^2\mathbf{U}\mathbf{U}^T + p(1-p)\delta(\mathbf{U}\mathbf{U}^T). \end{aligned} \quad (7.5)$$

Here $\delta(\cdot)$ resets all entries of the matrix to zeros except the diagonal elements.

Each M_i corresponds to a unique AU that is highly related to a specific geometric location. The filters are very likely to be sparse. In addition to the standard ℓ_2 regularizer $\|\mathbf{M}\|_2^2$, we thus put $\|\mathbf{M}\|_1$ to the optimization target to encourage the sparsity.

Upon integrating these specified constraints, we obtain the following optimization function to be minimized

$$\begin{aligned} [\mathbf{M}, \mathbf{B}] = \arg \min_{\mathbf{M}, \mathbf{B}} & \frac{1}{N} \sum_{i=1}^N \|X_i^T \mathbf{M} - U_i^T \mathbf{B}\|^2 + \\ & \alpha \|\mathbf{M}\|_2^2 + \beta \|\mathbf{M}\|_1 + \gamma r(\mathbf{B}), \end{aligned} \quad (7.6)$$

where α , β and γ are weighted factors. Figure 7.2 illustrates all terms in the optimization target. In this manner, the original AU recognition problem is effectively changed to a sparse tagging-like one, which is experimentally justified to have competitive performance but remarkably faster computational speed than other state-of-the-art algorithms.

7.3.2 Optimization

For the optimization target above, we use a block-coordinate descent scheme [11]. Fixing one of \mathbf{B} and \mathbf{M} , the optimal value of the other is obtained by setting the derivative to 0.

When \mathbf{B} is given, concretely, the matrix \mathbf{M} is optimized as

$$\mathbf{M} = (\mathbf{B}\mathbf{U}\mathbf{X}^T - \beta)(\mathbf{X}\mathbf{X}^T + N\alpha I). \quad (7.7)$$

When \mathbf{M} is fixed, equivalently, the optimal \mathbf{B} is optimized by

$$\mathbf{B} = (\gamma \mathbf{S} + \mathbf{M}\mathbf{X}\mathbf{U}^T)(\gamma \mathbf{T} + \mathbf{U}\mathbf{U}^T)^{-1}. \quad (7.8)$$

We iteratively optimize \mathbf{M} and \mathbf{B} until either the iteration number reaches maximum or the following criterion is satisfied:

Algorithm 7.1 Training process

Require:

Expression Set: \mathbf{X}
Action Unit of \mathbf{X} : \mathbf{U}
Maximum Iteration for Optimization: MaxIter

Output:

Filter Set: \mathbf{M}
1: Initiate \mathbf{B} randomly
2: **repeat**
3: Calculate \mathbf{M} using Eq. (7.7)
4: Calculate \mathbf{B} using Eq. (7.8)
5: **if** \mathbf{M} satisfies Eq. (7.9) **then**
6: **break**
7: **else**
8: MaxIter := MaxIter - 1
9: **end if**
10: **until** MaxIter = 0
11: **return** \mathbf{M}

$$\frac{||\mathbf{M}_{(k)} - \mathbf{M}_{(k-1)}||_F}{||\mathbf{M}_{(k-1)}||_F} < \epsilon \text{ and } \frac{||\mathbf{B}_{(k)} - \mathbf{B}_{(k-1)}||_F}{||\mathbf{B}_{(k-1)}||_F} < \epsilon , \quad (7.9)$$

where $\mathbf{M}_{(k)}$ and $\mathbf{B}_{(k)}$ are the value of \mathbf{M} and \mathbf{B} of the k -th round respectively, and ϵ is a small positive value. The $|| \cdot ||_F$ is the Frobenius norm.

For better understanding, a pseudocode is stated in Algorithm 7.1.

The proposed filter M_i has a good geometrical interpretation. The reason is that if we reshape it in accordance with the input image patch, the filter will naturally span a mask for the particular AU. That is, the element at M_i serves as a weight for the corresponding pixel in the input expression image. For example, in Fig. 7.1, the second filter corresponds to the AU of *Cheek Raiser*. Therefore, we expect the most distinguishable part to be in the area of the cheek, as illustrated in the red boxes on the filter.

7.3.3 Joint Learning of Action Units and Facial Expressions

After encoding the model in matrix \mathbf{M} , we can process some unseen expressions. As in Fig. 7.3, for a new expression image with size $m \times n$, we vectorize it as $X_{new} \in \mathbb{R}^{mn \times 1}$, followed by applying the filters as

$$R = X_{new}^T \mathbf{M} . \quad (7.10)$$

Clearly, the response R will be a vector of length L , and L is the number of AUs of interest. Each element of R indicates the possibility of corresponding AU existing in the input expression. Note that the elements of R may vary beyond the scope of

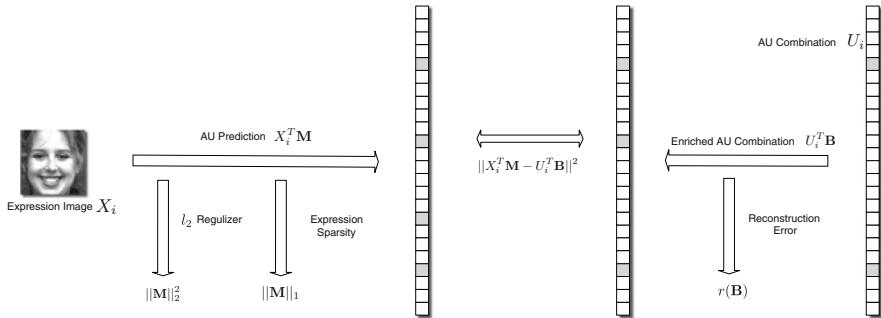


Fig. 7.3 The recognition process. On the obtained raw prediction of AU combination, a normalization process is executed to map the prediction to [0, 1]. The normalized prediction is thresholded to obtain the final AU prediction. An additional linear classifier is trained on the normalized prediction for the estimate of the emotional state

Algorithm 7.2 Action unit and emotion recognition

Require:

Learned Filter Set: \mathbf{M}

Learned Coefficient Matrix: θ

Input Expression: X_{new}

Output:

AU Combination Prediction \hat{U}

Emotion State Prediction $emotion$

1: $R := X_{\text{new}}^T \mathbf{M}$ {Apply Learned Model on Input Expression}

2: $R_{\text{norm}} := \frac{R - \min(R)}{\max(R) - \min(R)}$
 {Normalize responses to [0, 1]}

3: $idx := \text{find}(R_{\text{norm}} >= 0.5)$
 {find an index of normalized response which are larger than 0.5}

4: $\hat{U}[idx] := 1, \hat{U}[\sim idx] := 0$
 {normalized response larger than 0.5 indicates the activation of corresponding AU}

5: $E := \theta R_{\text{norm}}$

6: $emotion = \arg \max E$
 {select the emotion with the strongest response}

7: **return** $\hat{U}, emotion$

[0, 1]. Therefore, the R is normalized and thresholded to {0, 1} to get the prediction, as the vector shown in the upper right of Fig. 7.3. The whole procedure for AU recognition is described in Algorithm 7.2.

Once the AU combination prediction is completed, a classifier is employed to jointly select the most likely emotion state based on the normalized prediction values, as shown in Fig. 7.3. For simplicity, we assume a linear relationship between the AU prediction and the emotional state.

Specifically, let E_i be the 1-of-K coding of the emotional state of i -th expression sample. That is, $E_{ik} = 1$ and $E_{ij} = 0$ for all $j \neq k$ if the i -th expression sample falls into the emotion class k . And \hat{U}_i is the normalized AU combination prediction.

Formally, the linear relationship between emotion state E_i and normalized AU combination prediction R_{norm} is written as

$$R_{\text{norm}-i}\theta = E_i . \quad (7.11)$$

Let $\mathbf{E} = [E_1, E_2, \dots]$ and $\mathbf{R} = [R_{\text{norm}-1}, R_{\text{norm}-2}, \dots]$, the over-determined problem can be estimated by the least square method

$$\theta = (\mathbf{R}^T \mathbf{R})^{-1} \mathbf{R}^T \mathbf{E} . \quad (7.12)$$

Given the estimated θ and a normalized AU prediction, inferring the emotion state simply requires multiplying the normalized AU prediction by θ and selecting the index of the strongest response. The procedure is included in Algorithm 7.2.

7.3.4 Efficiency Advantage

The main advantage of our method is its high speed. Traditional approaches are usually composed of two parts: feature representation and AU recognition. Our approach does not explicitly follow these steps. Instead, matrix multiplication is used to assign different weights for all pixels, which can be viewed as an implicit feature extraction process. Following this, the AU prediction is attained through the normalization and thresholding the matrix product.

To analytically show our speed advantage, Table 7.1 summarizes the number of main operations required by some representative feature extraction methods as well as our method.

Let's assume an input facial expression image has dimensions $m \times n$ and a total of L AUs. Calculation of the LBP image requires a comparison of every pixel to its (typically 8) neighbors; DCT-based approaches calculate the discrete cosine transformation coefficients and keep some low-frequency part, which involves mn cosine calculation; Gabor-based approaches calculate Gabor images with γ orientations and λ spatial frequencies, which costs $\gamma\lambda mn$ exponential computation.

Table 7.1 The number of operations for each method

Approaches	Operations
LBP [3]	$8nm$ comparison L classification
DCT [29]	nm cosine computation L classification
Gabor [18]	$O(\kappa L)$ exponential computation L classification
Our method	L vector multiplication 1 normalization and thresholding

However, after feature selection, κ elements for each AU are chosen. Thus, during testing, only these elements need calculation. These will be slightly less than κL as the selected elements may intersect. In addition, these methods require L separate classification operations for L AUs.

In our method, only one matrix multiplication is required, which can be viewed as L vector multiplications. The post-processing involves normalization and thresholding of the matrix product. All these steps can be executed in a shorter time relative to other methods.

7.4 Experiments

We conducted comprehensive experiments on four benchmark facial expression datasets, as well as a combined dataset, to evaluate the performance of our proposed algorithm. First, the four benchmark datasets and experiments protocol are introduced in Sect. 7.4.1. The prediction performance is compared in Sect. 7.4.2. Section 7.4.3 analyzes the validity of the sparsity assumption. Lastly, the computational speed of various algorithms is compared in Sect. 7.4.4.

7.4.1 Experimental Setup

The five benchmark expression datasets we have used consist of two grayscale image datasets and one depth image dataset. They are the CK+, FERA, BOSPHORUS, and Youtube face datasets.

The **CK+** (Cohn-Kanade+) is a grayscale facial expression dataset labeled with both AU and emotion [20]. This dataset consists of 10,708 expression images from 593 sequences that were collected from a total of 123 subjects. Because only one frame in each sequence is labeled with AUs to represent the whole sequence, these clearly labeled images are used for our experiments, resulting in a subset of 593 images. Figure 7.4 shows some example grayscale facial expression images from the CK+ dataset. Due to the dataset's limited scale, the leave-one-out scheme is used for cross-validation. For other datasets, a 10-fold cross-validation is employed for performance evaluation. It should be noted that we only report the emotion accuracy for this database in this chapter because it is a unique database labeled by emotional states.

The **FERA** (Facial Expression Recognition and Analysis) dataset, a grayscale image collection, was initially developed for an expression recognition competition [28]. It has predefined training and testing sets. Because the ground-truth labels of the testing set cannot be accessed, only the predefined training set is used for both training and testing in our experiment. The subset we used contains 66 clips of video sequences, totaling 3900 frames. Images from this dataset cannot be displayed due to the license agreement.



Fig. 7.4 Example grayscale images used for AU recognition. Cropped images from the CK+ dataset are shown

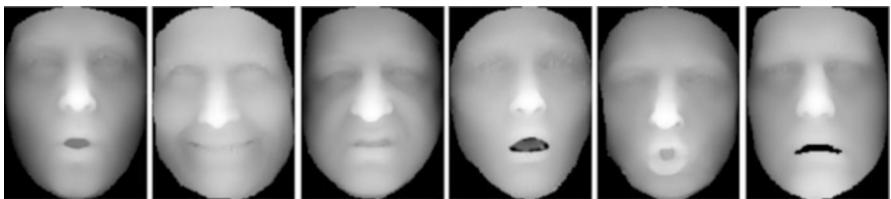


Fig. 7.5 The depth images from BOSPHORUS 3D dataset are offered in the scanner output format, providing 3D coordinates of points in order

We also construct a dataset, named **FERA & CK+**, by mixing all samples from CK+ and FERA. The purpose of the dataset is to test the robustness of different approaches under heterogeneous data distribution.

For these three image datasets, a Viola-Jones face detector [30] is applied to the first frame of each clip, followed by a tracker to crop all face images to a 200×200 -pixels bounding box.

The **BOSPHORUS** dataset [24, 25] provides expression data in the form of depth images. It contains 4666 captures of facial expressions, each composed of both a color image and a depth image. In our experiments, only depth images of unoccluded frontal faces were used, resulting in a subset of 2690 depth images for both training and testing. Figure 7.5 shows some examples from the BOSPHORUS datasets. The depth data is normalized to the $[0, 1]$ range and used as the gray level for visualization. Areas appearing brighter are closer to the camera relative to darker areas.

The **Youtube Faces Database** contains 3425 videos featuring 1595 individuals [33], and a total of 621,126 frames without expression labels are involved. We will utilize our algorithm to assign expression labels to this real-world multimedia dataset.

We employ two metrics, precision and recall, to evaluate the performance of different algorithms. *Precision* quantifies the fraction of detected AUs that genuinely

exist in the expression; *recall* quantifies the fraction of AUs present in the expression that the algorithm has accurately detected. These metrics are defined in Eq. (7.13), where TP, FP, and FN denote true positive, false positive, and false negative, respectively. The term ‘positive’ denotes the presence of an AU, and ‘negative’ the absence. We use the *F1-score* as our primary measure, representing the harmonic mean of precision and recall.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (7.13)$$

For comparison, we consider several potential algorithms, as discussed in Sect. 7.2. We have implemented recent approaches as reported in [3, 18, 29]. Among these, the Gabor-based approach [18] achieves the best performance. Hence, for simplicity, we confine our comparison to this method. In the Gabor-based approach, each cropped input image is resized to dimensions of 96×96 pixels and convolved with a bank of 72 Gabor filters, which consist of 8 orientations and 9 spatial frequencies, resulting in a 663,552-dimensional feature vector. As in [1], we employ Adaboost for feature selection, resulting in 20 single dimensions being selected for each AU. The ultimate feature vector is a union of all selected features. The classifier employed is an SVM with the RBF kernel. All features are normalized prior to classification. It should be noted that the optimal parameter combination is chosen based on grid search.

Apart from the Gabor-based approach, we present two other alternatives, an LBP-based approach [3] and a DCT-based approach [29], for the comparison of recognition speed.

The computational platform used in our experiments has four AMD Opteron 6378 processors, each with 16 cores. The platform has a total of 512 GB of RAM. The software environment includes Matlab R2013b running on Debian 3.2. For the time consumption experiment, only one core is used.

7.4.2 Performance Comparison

In this section, we compare the performance of several algorithms, including our own, in recognizing AUs and emotional expression.

7.4.2.1 AU Recognition

Table 7.2 presents the F1-scores for various AUs and the average F1-score across four datasets. On the CK+ and FERA datasets, our algorithm slightly underperforms

Table 7.2 Recognition accuracy comparison on the CK+, FERA, FERA & CK+, and BOSPHORUS datasets. Upper left: CK+; upper right: FERA; lower left: FERA & CK+; lower right: BOSPHORUS. The results are measured in F1-scores

AU #	Ours	Gabor	AU #	Ours	Gabor
1	0.65	0.67	1	0.86	0.83
2	0.59	0.67	2	0.83	0.73
4	0.60	0.60	4	0.91	0.79
5	0.53	0.49	6	0.94	0.87
6	0.52	0.54	7	0.87	0.87
7	0.45	0.48	10	0.88	0.81
9	0.56	0.60	12	0.91	0.64
12	0.65	0.63	15	0.81	0.92
15	0.44	0.55	17	0.63	0.77
17	0.70	0.70	18	0.60	0.83
25	0.81	0.76	25	0.57	0.72
27	0.61	0.81	26	0.55	0.72
Mean	0.593	0.625	Mean	0.780	0.792

AU #	Ours	Gabor	AU #	Ours	Gabor
1	0.77	0.84	1	0.26	0.21
2	0.74	0.78	2	0.24	0.18
4	0.79	0.77	4	0.26	0.21
6	0.85	0.86	5	0.23	0.21
7	0.82	0.83	7	0.52	0.60
10	0.78	0.80	12	0.30	0.26
12	0.85	0.72	17	0.22	0.20
15	0.70	0.71	25	0.46	0.40
17	0.58	0.54	26	0.25	0.22
25	0.57	0.68	Mean	0.304	0.277
Mean	0.745	0.753			

the Gabor-based method with regard to the mean F1-score, scoring 0.032 lower on CK+ and 0.012 lower on FERA, respectively. Despite this, the results indicate that our algorithm maintains comparable accuracy with the leading approach, with the added advantage of significantly faster computational speed (to be revealed later).

Both methods achieve better performance on FERA than on CK+. This could be attributed to the larger sample size of FERA, providing more training data for the model. Additionally, the FERA images are sourced from video sequences, resulting in high similarity between many of them, which can potentially enhance accuracy.

Concerning the FERA & CK+ dataset, the results of both methods are closer to FERA than to CK+, and the performance difference between our algorithm and the Gabor-based one narrows further. This is likely due to (1) FERA accounts for the most part of the dataset, hence exerting more influence, and (2) our algorithm has better robustness to this heterogeneous dataset than the Gabor-based algorithm.



Fig. 7.6 Example learned filters on CK+ dataset. Note the resemblance between mask images and corresponding AUs. AUs from top to bottom, from left to right: 13: Sharp Lip Puller; 22: Lip Funneler; 29: Jaw Thrust; 30: Jaw Sideways; 34: Cheek Puff; 44: Eyebrow Gatherer; 54: Head Down; 61: Eyes Turn Left

It is worth mentioning that our algorithm outperforms the Gabor-based approach in most cases for the BOSPHORUS dataset, indicating the adaptability of our algorithm to both grayscale images and depth images.

As mentioned in Sect. 7.3, the learned matrix \mathbf{M} assigns weight to expression images and serves as masks for AUs of interest. Figure 7.6 shows some examples of learned filters on CK+. The relationship between learned filters and their corresponding AUs is illustrated in the mask images. For instance, the image in the top-left of the figure shows a flat mouth, indicative of a person pulling their lip—an action which corresponds to the AU *Sharp Lip Puller*. This correspondence between AUs and mask images is visible across other examples as well.

Besides, all learned mask images shown in Fig. 7.6 bear a resemblance to human faces. However, this phenomenon may not be desirable as it suggests that areas outside of the decision region (e.g., the mouth area in the top-left of Fig. 7.6) of the face may serve as noise for later classification. If a model is shown a great resemblance to the top-left image of Fig. 7.6, but with an open mouth, our model may recognize it as positive for AU 13 because other areas match well.

This issue could be alleviated in a large-scale dataset, as evident in the learned filters of the FERA dataset in Fig. 7.7. It is observable that, in contrast to the learned filters of the CK+ dataset shown in Fig. 7.6, these filters less closely resemble human faces. For instance, in the case of AU 26, even though the facial structure is barely discernible, the region corresponding to the jaw (at the lower center of the mask image) has a different pattern compared to other regions. This may stem from the use of larger and thus sufficient training data. Consequently, the model generalizes better because it does not resemble any specific individuals.

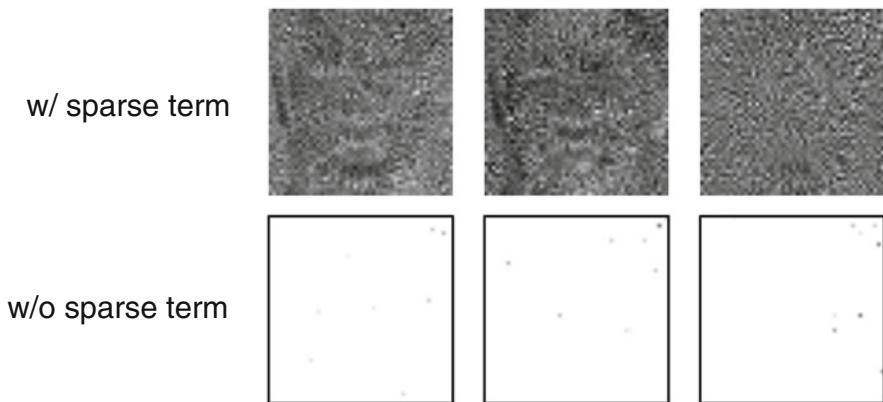


Fig. 7.7 Example learned filters on the FERA dataset. The first row is learned with the sparse term. The second row is learned without sparse term. AUs from left to right: 7: Lid Tightener; 12: Lip Corner Puller; 26: Jaw Drop

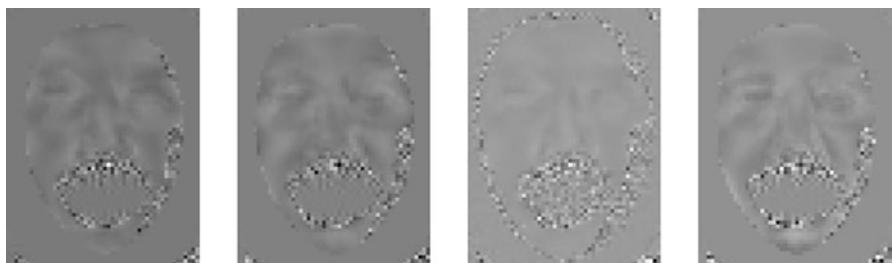


Fig. 7.8 Example learned filters on the BOSPHORUS Dataset. AUs from left to right: 4: Brow Lowerer; 12: Lip Corner Puller; 18: Lip Pucker

As for the BOSPHORUS dataset in Fig. 7.8, the resemblance between mask images and corresponding AUs are not straightforward. Nonetheless, we can find some clues: for example, the mouth in the third column has a hole, which represents the shape of *Lip Pucker*. This might be attributed to the depth images being not a reliable medium to tell expressions. Furthermore, the background areas appear uniformly gray because these areas are preprocessed and removed as non-face areas in the input data. As a result, patterns associated with emotions tend to concentrate around the mouth area. This observation is consistent with the intuition, i.e., the mouth is the most distinguishable area in depth facial images, as shown in Fig. 7.5.

7.4.2.2 Emotion Recognition

Given the predicted AU combination, we train a classifier to predict seven emotional states. The distribution of these emotions is detailed in Table 7.3. As mentioned

Table 7.3 Numbers of emotion samples in CK+ dataset

Emotion state	Number of instances
Anger	45
Contempt	18
Disgust	59
Fear	25
Happy	69
Sad	28
Surprise	83

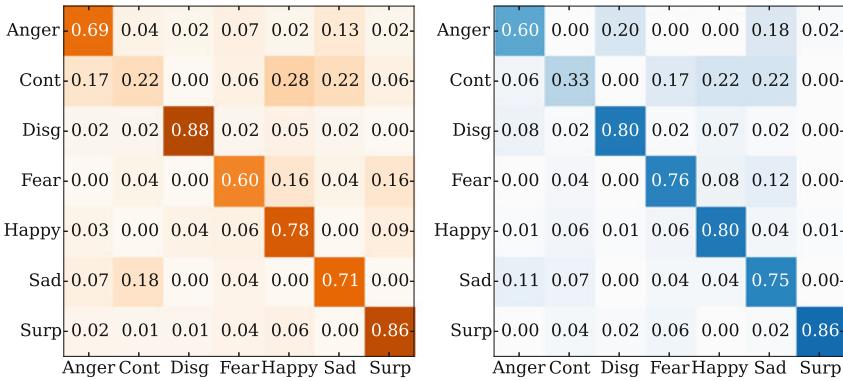


Fig. 7.9 Confusion matrices of emotion recognition on the CK+ dataset. Left: Our approach; right: Gabor-based approach

above, our method uses the normalized AU prediction with a linear classifier. To be fair, the Gabor-based approach uses the same algorithm to train its classifier. And the input data come from the output of AU prediction with discrete values.

Figure 7.9 shows the confusion matrices for emotion classification. The values on the diagonal are much deeper in color than other values, indicating high classification performance. The overall accuracy of our method is 75.23%, while the Gabor-based approach achieves 75.54%. The differences between our approach and the Gabor-based one are subtler in emotion recognition than in AU recognition. This is because our algorithm models the AU combination in a joint fashion. In this way, it can offer a more meaningful AU combination, and thus aids in more accurate emotion recognition.

We observe that for contempt emotion, the classification accuracy is relatively low for either method. This could potentially be due to the insufficient quantity of training samples. As shown in Table 7.3, the contempt class possesses the least number of samples within the dataset. Meanwhile, contempt expressions vary substantially because it usually involves asymmetric nose wrinkles that can appear on either side. This underscores the need for more training samples to enhance the generalization capabilities.

Table 7.4 F1-score of our method with and without the sparse term on the FERA dataset

AU #	With sparse term	Without sparse term
1	0.86	0.82
2	0.83	0.78
4	0.91	0.84
6	0.94	0.88
7	0.87	0.82
10	0.88	0.84
12	0.91	0.87
15	0.81	0.75
17	0.63	0.59
18	0.60	0.55
25	0.57	0.55
26	0.55	0.50
Mean	0.780	0.734

7.4.3 Effect of Model Choice

To validate our sparsity hypotheses on Action Units, we evaluate a model that does not contain the sparse term in the optimization function. As shown in Table 7.4, this alternative model is much less accurate in predicting AUs.

Figure 7.7 shows some example learned filters on the FERA dataset. The first row shows mask images learned with the sparse term. The second row shows mask images learned without the sparse term. It is evident that the latter is composed of an almost entirely white area, punctuated by sporadic darker dots, indicating that the learned filters are composed of very high values. That is, the learned filters are dense. In addition, because most positions in these filters share similar weights (white), they help little in distinguishing different AUs. It justifies the importance of the sparsity constraint we introduced in our proposed algorithm.

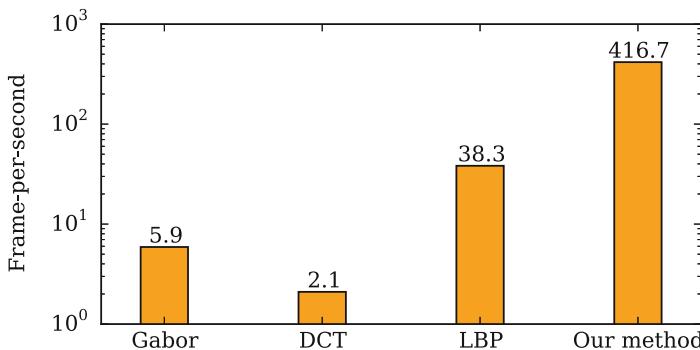
7.4.4 Computational Speed

We used 1000 input images from the FERA dataset to evaluate the time performance of different approaches. We set the dimensions of the input images to 200×200 . All images are preloaded into memory to avoid additional I/O operations.

Table 7.5 provides a comparison of the runtime for different approaches. The SVM utilizes an RBF kernel. The difference in SVM classification mainly originates from variations in feature length. For the LBP, a mapping is used to transform binary patterns into a 59-D vector. In the case of DCT, the input image is further partitioned into 8×8 patches; within each patch, the type-II DCT is extracted, but only the first 10 components are preserved, resulting in a 640-D vector. For the Gabor, the features cannot be used directly due to their extremely high dimensions.

Table 7.5 Runtime for different visual approaches

Feature	Extraction	SVM	Total
LBP [3]	0.0093	0.0168	0.0261
DCT [29]	0.0212	0.4607	0.4819
Gabor [18]	0.0027	0.1664	0.1691
	Resizing	Model application	Total
Ours	0.0024	3.8×10^{-5}	0.0024

**Fig. 7.10** Framerate measure (per-second) of different methods

As a workaround, we select 20 common AUs and utilize Adaboost to choose 20 single dimensions for each AU. The union set of these selected features forms a 333-D vector, accounting for some intersection.

For a straightforward comparison, we count how many frames each method can process per second. Figure 7.10 shows the frame-per-second (fps) performance of each method, with our method demonstrating the highest speed.

Three key factors contribute to this performance advantage. Firstly, we do not explicitly calculate any visual features. Instead, our model performs a simple matrix multiplication operation, which is inherently fast. Secondly, our approach does not require any dimension reduction or feature selection processes, which are not negligible in the case of high-dimensional features such as DCT. Lastly, our classification process is implicitly incorporated within the model multiplication. Conversely, other methods employing SVM with a non-linear kernel require additional computational time to model each AU separately.

7.4.5 Tagging a Real-World Dataset

To further demonstrate the capacity of our algorithm to recognize facial expressions in real-world scenarios, we use the model, trained on the CK+ dataset, to generate emotion tags for the Youtube Faces Database [33].



Fig. 7.11 Tagging results of the Youtube Faces Database. Emotional states from the top row to the bottom row: anger, contempt, disgust, fear, happiness, sadness, and surprise. The state is displayed at the top of the green frame in each image (clearer after zooming in). The videos are sourced from YouTube, and all associated copyrights are retained by the original content creators. The still frames are incorporated in the figure for illustrative purposes and to support the conceptual discussions in this chapter. The authors acknowledge and appreciate the work of the original content creators

For each picture in the dataset, a Viola-Jones face detector [30] is applied to isolate the faces. Considering the face sizes in the dataset, we set the minimum size of the detected patch to 20×20 pixels. The cropped face patches are rescaled to the same size as our trained filters, i.e. 50×50 -pixel patches. We then apply the algorithm described in Algorithm 7.2 for both AU and emotional state recognition.

Figure 7.11 presents some tagging results grouped by emotional states. The results indicate that our method offers reliable tagging performance even when trained on a different dataset. However, we have observed some limitations: (1) Since our model is trained under frontal face images, non-frontal faces (see the last column of the first row) may be misclassified; (2) When applied to individuals of races not included in the training set (last column of the second row), the model may mis-classify. These issues could be mitigated by using larger-scale multi-view, and multi-race expression datasets, which exceeds the scope of this work.

7.5 Conclusion and Future Work

In this chapter, we have proposed a novel facial expression recognition approach by regarding AUs as image tags. Instead of learning independent binary classifiers on pre-extracted visual features for each AU, we jointly and simultaneously learn all filters in the original image space. Because only matrix multiplication is involved, our model delivers remarkably fast computational speed. Meanwhile, we introduced the sparsity constraint to improve the interpretability of our tagging-based approach. Experiments in four benchmark datasets indicated that our approach is robust, efficient, and interpretable, compared with state-of-the-art algorithms. We also showed the potential of our method to make emotional tags for real-world multimedia such as photos or videos.

A possible future direction is to exploit the relationship between AU filters for achieving higher recognition accuracy without degenerating the speed advantage of our proposed approach.

References

1. Bartlett, M.S., Littlewort, G.C., Frank, M.G., Lainscsek, C., Fasel, I.R., Movellan, J.R.: Automatic recognition of facial actions in spontaneous expressions. *J. Multimed.* **1**(6), 22–35 (2006)
2. Chen, M., Zheng, A., Weinberger, K.: Fast image tagging. In: Proceedings of the International Conference on Machine Learning (2013)
3. Chew, S.W., Lucey, P., Lucey, S., Saragih, J., Cohn, J.F., Sridharan, S.: Person-independent facial expression detection using constrained local models. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition (2011)
4. Chibelushi, C.C., Bourel, F.: Facial expression recognition: a brief tutorial overview. In: CVonline: On-Line Compendium of Computer Vision (2003)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: ideas, influences, and trends of the new age. *ACM Comput. Surv.* **40**(2), 1–60 (2008)
6. Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Stanford University, Palo Alto (1978)
7. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. *Pattern Recogn.* **36**(1), 259–275 (2003)
8. Feng, Z., Feng, S., Jin, R., Jain, A.K.: Image tag completion by noisy matrix recovery. In: Proceedings of the European Conference on Computer Vision (2014)
9. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**(1), 119–139 (1997)
10. Garcia, I., Bronte, S., Bergasa, L.M., Almazan, J., Yebes, J.: Vision-based drowsiness detector for real driving conditions. In: IEEE Intelligent Vehicles Symposium (2012)
11. Hsieh, C.J., Chang, K.W., Lin, C.J., Keerthi, S.S., Sundararajan, S.: A dual coordinate descent method for large-scale linear SVM. In: Proceedings of the International Conference on Machine Learning (2008)
12. Ilbeygi, M., Shah-Hosseini, H.: A novel fuzzy facial expression recognition system based on facial feature extraction from color face images. *Eng. Appl. Artif. Intel.* **25**(1), 130–146 (2012)
13. Jeni, L.A., Girard, J.M., Cohn, J.F., De La Torre, F.: Continuous AU intensity estimation using localized, sparse facial feature space. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition (2013)

14. Kumano, S., Otsuka, K., Mikami, D., Yamato, J.: Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings. In: Proceedings of the International Conference on Multimodal Interaction (2009)
15. Lajevardi, S.M., Hussain, Z.M.: Automatic facial expression recognition: feature extraction and selection. *Signal Image Video Process.* **6**(1), 159–169 (2012)
16. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**(6755), 788–791 (1999)
17. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(9), 1075–1088 (2003)
18. Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., Bartlett, M.: The computer expression recognition toolbox (CERT). In: Proceedings of the International Conference on Automatic Face and Gesture Recognition (2011)
19. Long, F., Wu, T., Movellan, J.R., Bartlett, M.S., Littlewort, G.: Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. *Neurocomputing* **93**, 126–132 (2012)
20. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of the International Conference on Computer Vision Workshops (2010)
21. Porter, S., Brinke, L.t.: Reading between the lies: identifying concealed and falsified emotions in universal facial expressions. *Psychol. Sci.* **19**(5), 508–514 (2008)
22. Reeb-Sutherland, B.C., Rankin Williams, L., Degnan, K.A., Pérez-Edgar, K., Chronis-Tuscano, A., Leibenluft, E., Pine, D.S., Pollak, S.D., Fox, N.A.: Identification of emotional facial expressions among behaviorally inhibited adolescents with lifetime anxiety disorders. *Cognit. Emot.* **29**(2), 372–382 (2015)
23. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: Proceedings of the International Conference on Computer Vision (2009)
24. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: Proceedings of the First European Workshop on Biometrics and Identity Management, pp. 47–56. Springer, Berlin (2008)
25. Savran, A., Sankur, B., Bilge, M.T.: Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units. *Pattern Recogn.* **45**(2), 767–782 (2012)
26. Shreve, M., Godavarthy, S., Goldgof, D., Sarkar, S.: Macro-and micro-expression spotting in long videos using spatio-temporal strain. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition (2011)
27. Thiam, P., Meudt, S., Kächele, M., Palm, G., Schwenker, F.: Detection of emotional events utilizing support vector methods in an active learning HCI scenario. In: Proceedings of the Emotion Recognition in the Wild Challenge and Workshop (2014)
28. Valstar, M.F., Jiang, B., Mehu, M., Pantic, M., Scherer, K.: The first facial expression recognition and analysis challenge. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition (2011)
29. Velusamy, S., Kannan, H., Anand, B., Sharma, A., Navathe, B.: A method to infer emotions from facial action units. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (2011)
30. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (2001)
31. Vonikakis, V., Winkler, S.: Emotion-based sequence of family photos. In: Proceedings of the ACM Multimedia Conference (2012)
32. Wang, S.J., Chen, H.L., Yan, W.J., Chen, Y.H., Fu, X.: Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. *Neural Process. Lett.* **39**(1), 25–43 (2014)
33. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (2011)
34. Zhang, L., Tjondronegoro, D.: Facial expression recognition using facial movement features. *IEEE Trans. Affect. Comput.* **2**(4), 219–229 (2011)

Chapter 8

ExpressionFlow: A Microexpression Descriptor for Efficient Recognition



Feng Xu, Yifan Yuan, Junping Zhang, and James Z. Wang

Abstract Microexpressions are involuntary facial movements that often reflect a person’s true emotions. Their fleeting nature and subtle shifts, however, make them challenging to detect. Our earlier work, the Facial Dynamics Map, represented a microexpression by estimating dense optical flow. Although it achieved high prediction accuracy, it was inefficient in feature extraction and lacked magnitude information. In this chapter, we address these issues by proposing ExpressionFlow, a novel descriptor which directly captures the dominant motion patterns in microexpression image sequences. Geometrically intuitive and relatively easy to implement, ExpressionFlow reflects the nature of microexpressions while preserving complete information. Comparative experiments on four benchmark datasets suggest that our method attains the best performance in real-time compared with other state-of-the-art algorithms.

8.1 Introduction

Facial expressions serve as immediate manifestations of human emotions. Over the past decade, automatic facial expression recognition has emerged as an active area

F. Xu

School of Computer Science, Fudan University, Shanghai, China
Ant Financial Group, Hangzhou, China
e-mail: feng_xu@fudan.edu.cn

Y. Yuan · J. Zhang (✉)

School of Computer Science, Fudan University, Shanghai, China
e-mail: yfyan21@m.fudan.edu.cn; jpzhang@fudan.edu.cn

J. Z. Wang

Data Science and Artificial Intelligence Area, and Human-Computer Interaction Area, College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA
e-mail: jwang@ist.psu.edu

of research in computer vision and affective computing [4, 15, 45]. The algorithms for detecting facial expressions are described in the previous chapter of this volume.

Microexpressions, in contrast to macroexpressions, represent a special form of subtle and fleeting spontaneous facial activity that often conceals a person's genuine emotions. Because microexpressions cannot be suppressed or feigned, they offer a critical element in affective estimation with broad implications in public safety and psychological therapy [1, 11, 20, 24, 25, 36].

Macroexpressions typically last from 3/4 seconds to 2 seconds, while microexpressions only last for 1/25 to 1/3 of a second [39]. Besides, while macroexpressions can occur over single or multiple facial regions depending on the category of expression, microexpressions never manifest simultaneously on the upper and lower parts of the face [23, 39] and their amplitude is relatively small. Therefore, identifying microexpressions poses a far more complex research challenge than their macro counterparts.

Compared with macroexpression studies, the history of microexpression discovery is shorter. In 1966, Haggard et al. [13] first proposed the idea of microexpression. Since then, Ekman et al. [8] reported a psychological case related to microexpression. In a conversation between a psychologist and depression patients [8], patients who often smile have several frames of very painful expressions. The researchers regarded the rapid, unconscious, and spontaneous facial movements produced by people when experiencing strong emotions as microexpressions.

Microexpressions are highly reliable in emotion recognition tasks and have potential applications in areas such as marital relationship prediction [12], communication negotiation [26], and teaching evaluation [7, 36]. In addition to sentiment analysis, researchers have observed microexpressions generated when an individual is intentionally lying [23]. Besides, it has been found that microexpression recognition training can enhance lie detection skills [35].

After being discovered, microexpression has attracted the attention of the research community [9, 10, 24, 30, 42]. However, the automatic recognition of microexpressions has only been explored in recent years [21, 32, 38]. It is highly desirable to develop an automatic microexpression recognition system.

Several approaches have been proposed to tackle this problem. Polikovsky et al. [22] utilized a 3D orientation gradient histogram to describe microexpression in the spatiotemporal domain. Shreve et al. [28, 29] calculated a strain map based on dense optical flow fields to indicate the deformation status of the facial area. Both studies used posed datasets for validation, potentially introducing disingenuous data samples.

Pfister et al. [19, 21] utilized a spatiotemporal local texture descriptor, LBP-TOP (Local Binary Pattern in Three Orthogonal Planes) [44], to characterize a microexpression over time. Wang et al. [32] improved this work by using an independent color space for better feature representation. In another work, Wang et al. [33] modeled the face as a 3-order tensor, and then applied discriminant tensor subspace analysis (DTSA) for feature learning, and used a neural network for pattern classification.

Despite these advances, a notable shortcoming in the existing literature is the absence of an intuitive relationship between the proposed methods and the nature of microexpressions. In other words, most existing approaches operate in a “black box” fashion where the extracted features cannot provide intuition for understanding microexpressions. In our prior work, we introduced the Facial Dynamics Map (FDM) to characterize face movement for microexpression recognition [38], which demonstrated its effectiveness on benchmark datasets.

However, a key limitation of FDM lies in its need to extract dense optical flow fields, which is a time-consuming process and the cost is too high for microexpression which only lasts a short time, leading to the difficulty of real-time detection. Further, in the iterative procedure to calculate the principal motion direction, the motion magnitude may be lost.

To address these issues while preserving the benefits of FDM, we propose an efficient descriptor for microexpression recognition in this chapter. First, an expression video clip is divided into small cuboids. For each cuboid, a 2-dimensional motion vector is calculated to capture the spatiotemporal dynamics. By unifying these motion vectors, we obtain the ExpressionFlow descriptor for each microexpression sequence. The ExpressionFlow is used as a feature representation for microexpressions, and a common classifier is employed to recognize the affective state of a given subject.

The main **contributions** of our work are as follows:

1. We propose a descriptor for microexpression that preserves both motion direction and magnitude. The recognition results achieved with this descriptor surpass those obtained with existing methods.
2. The proposed descriptor is computationally efficient. Our method is capable of processing microexpressions in real-time.
3. The proposed descriptor directly relates to the motion patterns of microexpressions. By analyzing the descriptor, the actual facial activity associated with a microexpression can be visualized.

The remainder of the chapter is organized as follows. Section 8.2 reviews the related work in microexpression recognition. Section 8.3 describes our novel descriptor, ExpressionFlow, and the entire procedure for microexpression recognition. Experimental results are presented and analyzed in Sect. 8.4. Finally, we conclude and discuss future directions in Sect. 8.5.

8.2 Related Work

The existing literature presents two primary strategies for recognition: the local strategy and the holistic strategy. To be specific, approaches that collect clues in sub-regions of the face and make determinations fall into the former category. These sub-regions are often split in accordance to facial elements, such as mouth, cheeks, nose, etc. In the latter case, the face is regarded as an entire unit for analysis.

Under the local strategy, facial sub-regions are often determined by facial landmark models such as Active Shape Model (ASM) [5], Active Appearance Model (AAM) [6], or Constrained Local Model (CLM) [27]. These models locate the facial landmarks as trained, and thus can be utilized to segment a face. For example, Polikovsky et al. [22] employed an ASM to segment the face into 12 sub-regions. Within each sub-region, the spatiotemporal derivatives in three coordinates are extracted to describe the skin movement. Afterward, a final vote is taken for Action Unit recognition. Shreve et al. [28, 29] similarly used an ASM model for face segmentation, after which they estimated dense optical flow fields to calculate strain maps. Indicating the elastic modulus of each point in the face, the train maps can tell the level of the tenseness of facial actions.

However, local methods, heavily reliant on prior knowledge of human faces, can compromise microexpression recognition performance due to potential inaccuracies in face segmentation.

On the other hand, holistic methods view the face in its entirety and determine the emotional state of the subject. For instance, Pfister et al. [19, 21] used the LBP-TOP descriptor to represent each facial image sequence. The Temporal Interpolation Model (TIM) [46] is used to align the frame sequences to the same length. An aligned sequence is then decomposed into smaller spatiotemporal cuboids, within which the LBP-TOP feature is extracted and linked together. In this work, an ASM is employed for facial image alignment, instead of segmentation. Wu et al. [37] combined the Gabor feature descriptor and GentleSVM to recognize microexpressions. Because the Gabor operator produces high-dimensional features, additional feature selection or dimension reduction is necessary. Several attempts have been made to enhance these features' performance in both preprocessing stage and acquiring more complete information [14, 17, 33, 34, 43].

Approaching the problem from a different perspective, Wang et al. [32], instead of extracting vision features, regarded a microexpression sequence as 3-order $S \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. An optimal projection is learned to process new samples. In general, the algorithm is applicable to all video classification tasks and works with holistic approaches.

Before introducing our new method, we briefly review Facial Dynamics Map [38], which falls under the holistic category. In our previous work, dense optical flow fields are initially extracted between consecutive frames. These fields are viewed as spatiotemporal cuboid and further segmented into smaller cuboids. Within each smaller cuboid, we utilized an iterative algorithm to estimate its principal direction. These principal directions are quantized and linked as a feature vector for microexpression representation and recognition. While effective across several benchmark datasets, the method suffers from high computational costs, predominantly due to the dense optical flow field extraction, which involves slow numeric optimization.

To circumvent this issue, we explore the potential for direct microexpression dynamics estimation from image sequences.

8.3 Our Method

We now describe the proposed ExpressionFlow descriptor and the process for microexpression recognition. Microexpression recognition is further partitioned into two subtasks. *Microexpression identification* refers to the procedure of determining whether a facial clip contains a microexpression, irrespective of its emotional context, be it happiness or sadness. *Microexpression categorization* refers to the procedure of discerning the emotion state represented by a microexpression clip.

8.3.1 The ExpressionFlow

Our proposed ExpressionFlow aims to describe the dynamic pattern of facial action. Because the movements in a microexpression are subtle, it is necessary to leverage the following two properties of microexpression for accurate description:

- Facial surface moves in roughly the same spatial direction when the observation area is small enough. Meanwhile, a pixel-level movement description for microexpression is redundant.
- Facial surface moves in roughly the same temporal direction when the observation period is short enough, rendering a very high frame rate redundant.

Considering the two assumptions above, we aim to identify the motion pattern of each local spatiotemporal area.

Given a video sequence of dimension $X \times Y \times T$ (for a T -frame video, each frame is of size $X \times Y$), we equally split the images into $X \times Y$ patches. The dimension of each patch is $\lfloor \frac{X}{\bar{X}} \rfloor \times \lfloor \frac{Y}{\bar{Y}} \rfloor$, where $\lfloor a \rfloor$ is the largest integer not greater than a . In the temporal axis, we divide the whole video into T batches. This results in a set of cuboids of dimension $\lfloor \frac{X}{\bar{X}} \rfloor \times \lfloor \frac{Y}{\bar{Y}} \rfloor \times \lfloor \frac{T}{\bar{T}} \rfloor$. We define such a **segmentation scheme** as $[X, Y, T]$, where $X \times Y$ is the **space division**, and T is the **temporal division**.

For a microexpression sequence, specifically, we split the sequence into spatiotemporal cuboids. Thus, each cuboid contains consecutive image patches at the same location of the whole images, denoted as $I_1, I_2, \dots, I_{\frac{T}{\bar{T}}}$, each $I_t \in \mathbb{R}^{X \times Y}$.

We estimate a single movement vector (u, v) to represent the motion pattern of that spatiotemporal cuboid given the assumptions above. Ideally, a motion vector should satisfy the following equation:

$$I_t(x, y) = I_{t+1}(x + u, y + v), \quad (8.1)$$

where $t \in \{1, \dots, \lfloor \frac{T}{\bar{T}} \rfloor\}$, $x \in \{1, \dots, \lfloor \frac{X}{\bar{X}} \rfloor\}$, $y \in \{1, \dots, \lfloor \frac{Y}{\bar{Y}} \rfloor\}$.

If we set $C(x, y, t) = I_t(x, y)$, equivalently, we have:

$$C(x, y, t) = C(x + \Delta x, y + \Delta y, t + \Delta t). \quad (8.2)$$

Expanding the formula using Taylor series at the point (x, y, t) , we have

$$\begin{aligned} C(x, y, t) &= C(x, y, t) + \frac{\partial C}{\partial x} \Delta x + \frac{\partial C}{\partial y} \Delta y + \frac{\partial C}{\partial t} \Delta t \\ &\quad + \sum_{n=2}^{\inf} \left(\frac{1}{n!} \frac{\partial^n C}{\partial x^n} \Delta x^n + \frac{1}{n!} \frac{\partial^n C}{\partial y^n} \Delta y^n + \frac{1}{n!} \frac{\partial^n C}{\partial t^n} \Delta t^n \right). \end{aligned} \quad (8.3)$$

Ignoring the higher order terms, the above equation leads to

$$\frac{\partial C}{\partial x} \Delta x + \frac{\partial C}{\partial y} \Delta y + \frac{\partial C}{\partial t} \Delta t \approx 0. \quad (8.4)$$

In a camera-captured digital video clip, we set Δt to the temporal interval of two consecutive frames. Let $(u, v) = (\frac{\Delta x}{\Delta t}, \frac{\Delta y}{\Delta t})$, we have:

$$\frac{\partial C}{\partial x} u + \frac{\partial C}{\partial y} v = -\frac{\partial C}{\partial t}. \quad (8.5)$$

The equation should hold for all positions in the particular cuboid. That is,

$$\begin{aligned} \frac{\partial C}{\partial x}|_p u + \frac{\partial C}{\partial y}|_p v &= -\frac{\partial C}{\partial t}|_p, \\ p = (x, y, t) \in \{1, \dots, \lfloor \frac{X}{\Delta x} \rfloor\} \times \{1, \dots, \lfloor \frac{Y}{\Delta y} \rfloor\} \\ &\quad \times \{1, \dots, \lfloor \frac{T}{\Delta t} \rfloor\}. \end{aligned} \quad (8.6)$$

As long as the choices of (x, y, t) are more than 2, this is an over-determined problem and can be re-written as :

$$S\mathbf{u} = T, \quad (8.7)$$

where

$$S = \begin{bmatrix} \frac{\partial C}{\partial x}|_{1,1,1}, & \frac{\partial C}{\partial y}|_{1,1,1} \\ \frac{\partial C}{\partial x}|_{1,1,2}, & \frac{\partial C}{\partial y}|_{1,1,2} \\ \dots & \dots \\ \frac{\partial C}{\partial x}|_{\lfloor \frac{X}{\Delta x} \rfloor, \lfloor \frac{Y}{\Delta y} \rfloor, \lfloor \frac{T}{\Delta t} \rfloor}, & \frac{\partial C}{\partial y}|_{\lfloor \frac{X}{\Delta x} \rfloor, \lfloor \frac{Y}{\Delta y} \rfloor, \lfloor \frac{T}{\Delta t} \rfloor} \end{bmatrix}, \quad (8.8)$$

$$\mathbf{u} = [u, v]^T, \quad (8.9)$$

$$T = \begin{bmatrix} -\frac{\partial C}{\partial t}|_{1,1,1} \\ -\frac{\partial C}{\partial t}|_{1,1,2} \\ \dots \\ -\frac{\partial C}{\partial t}|_{\lfloor \frac{X}{X} \rfloor, \lfloor \frac{Y}{Y} \rfloor, \lfloor \frac{T}{T} \rfloor} \end{bmatrix}. \quad (8.10)$$

Therefore, the temporal derivatives T are factorized as the product of spatial derivatives S and motion vector \mathbf{u} .

8.3.2 The Optimization Process

If we target to minimize the ℓ_2 -norm error function:

$$E_{\ell_2} = \|S\mathbf{u} - T\|_2, \quad (8.11)$$

it is straightforward to solve for such \mathbf{u} . Taking the derivative of E_{ℓ_2} with respect to \mathbf{u} , we have:

$$\mathbf{u} = (S^T S)^{-1} S^T T, \quad (8.12)$$

where $(S^T S)^{-1} S^T$ is the Moore-Penrose pseudoinverse of S .

A shortcoming of ℓ_2 -norm is that it can be easily influenced by outliers. That is, a small noise in the image sequence can significantly contribute to the error function. Therefore, ℓ_1 -norm error function can to some extent alleviate the influence of outliers:

$$E_{\ell_1} = \|S\mathbf{u} - T\|_1. \quad (8.13)$$

To solve for such \mathbf{u} in ℓ_1 -norm error function, we utilize the Nelder-Mead simplex algorithm [18].

8.3.3 The Algorithm

A complete algorithm for extracting ExpressionFlow is shown in Algorithm 3 and is illustrated in Fig. 8.1. For a clip of microexpression, we first extract the ExpressionFlow descriptor. Then, for the extracted feature vector, an SVM classifier is trained to perform the detection or categorization of microexpressions.

The Nelder-Mead simplex algorithm in ExpressionFlow extraction is also contained in Algorithm 3. In the pseudo-code, we construct a simplex with three vertices v_1, v_2, v_3 . ϵ_E and ϵ_v are thresholds indicating a small enough simplex when

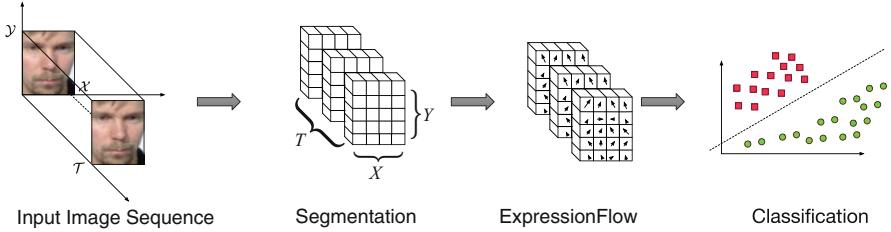


Fig. 8.1 Workflow of our approach. The segmentation scheme $[X, Y, T]$ is set to $[4, 4, 3]$ in this case

the optimization can terminate. And $\text{sort}(v_1, v_2, v_3, E)$ sorts the v_1, v_2, v_3 to the ascending order of $E(v_1), E(v_2), E(v_3)$.

8.3.4 Comparison to Facial Dynamics Map

Both FDM and ExpressionFlow are designed to create a compact representation of facial movements. FDM accomplishes this goal through processing extracted dense optical flow estimations, while ExpressionFlow directly calculates the representation from image sequences. ExpressionFlow offers several distinct advantages:

1. Despite the seemingly comprehensive information that dense optical flow provides, it can also contain significant noise. This issue may lead to accumulated noise impacting the outcome in a longer pipeline process.
2. To attain an accurate principal direction in FDM, we employed an iterative scheme that ignores motion magnitude. ExpressionFlow can effectively preserve the important motion magnitude.
3. ExpressionFlow can avoid a fine alignment procedure, which is mentioned in our previous work and is justified to have a subtle improvement in the aspect of effectiveness [38].

8.3.5 Computational Complexity

To recognize the fleeting microexpression, the speed is a critical index for different algorithms. Before justifying the empirical time consumption of different approaches in the experiments, we first analyze the theoretical complexity of each method.

We consider four candidates, i.e. LBP-TOP, DTSA, and ExpressionFlow with ℓ_1 -norm as well as ExpressionFlow with ℓ_2 -norm. Since all four methods share a common framework in computing the representation and recognizing microexpressions, we only compare their complexities in the feature representation stage.

Algorithm 3 ExpressionFlow**Require:**

Microexpression image sequence $I_1, I_2, \dots, I_T, I_t \in \mathbb{R}^{X \times Y}$
 segmentation scheme $[X, Y, T]$
 optimization type ℓ_1 or ℓ_2

Output:

ExpressionFlow feature vector

```

1: split the microexpression sequence into cuboids  $C_1, C_2, \dots, C_{X \times Y \times T}$ 
2: for all  $C_i$  do
3:   calculate spatial derivatives  $S_i$ 
4:   calculate temporal derivatives  $T_i$ 
5:   if optimization type =  $\ell_1$  then
6:     //optimize  $\mathbf{u}_i$  with Nelder-Mead simplex algorithm
7:      $\rho = 1, \chi = 2, \gamma = \frac{1}{2}, \sigma = \frac{1}{2}$ 
8:     randomly initiate a simplex with 3 vertices  $v_1, v_2, v_3$ 
9:     while  $|v_1 - v_2|_2 + |v_2 - v_3|_2 + |v_3 - v_1|_2 > \epsilon_v$  and  $|E(v_1) - E(v_2)|_2 + |E(v_2) - E(v_3)|_2 + |E(v_3) - E(v_1)|_2 > \epsilon_E$  do
10:     $v_1, v_2, v_3 = \text{sort}(v_1, v_2, v_3, E)$ 
11:     $\bar{v} = \frac{v_1 + v_2 + v_3}{3}$ 
12:     $v_r = \bar{x} + \rho(\bar{v} - v_3)$ 
13:    if  $E(v_1) \leq E(v_r) < E(v_3)$  then
14:       $v_1, v_2, v_3 = \text{sort}(v_1, v_2, v_3, E)$  continue
15:    else if  $E(v_r) < E(v_1)$  then
16:       $v_e = \bar{v} + \chi(v_r - \bar{v})$ 
17:      if  $E(v_e) < E(v_r)$  then
18:         $v_1, v_2, v_3 = \text{sort}(v_1, v_2, v_e, E)$  continue
19:      else if  $E(v_e) \geq E(v_r)$  then
20:         $v_1, v_2, v_3 = \text{sort}(v_1, v_2, v_r, E)$  continue
21:      end if
22:    else if  $E(v_r) \geq E(v_2)$  then
23:      if  $E(v_2) \leq E(v_r) < E(v_3)$  then
24:         $v_c = \bar{v} + \gamma(v_r - \bar{v})$ 
25:        if  $E(v_c) \leq E(v_r)$  then
26:           $v_1, v_2, v_3 = \text{sort}(v_1, v_2, v_c, E)$  continue
27:        end if
28:      else if  $E(v_r) \geq E(v_3)$  then
29:         $v_{cc} = \bar{v} - \gamma(\bar{v} - v_3)$ 
30:        if  $E(v_{cc}) < E(v_3)$  then
31:           $v_1, v_2, v_3 = \text{sort}(v_1, v_2, v_{cc}, E)$  continue
32:        end if
33:      end if
34:    end if
35:     $v_1, v_2, v_3 = v_1, v_1 + \sigma(v_2 - v_1), v_1 + \sigma(v_3 - v_1)$ 
36:  end while
37:   $\mathbf{u}_i = v_1$ 
38: end if
39: if optimization type =  $\ell_2$  then
40:    $\mathbf{u}_i = (S_i^T S_i)^{-1} S_i^T T_i$ 
41: end if
42: end for
43: ExpressionFlow =  $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{X \times Y \times T}\}$ 
44: return ExpressionFlow

```

As for ExpressionFlow, the algorithm involves computing motion vector \mathbf{u} for $X \times Y \times T$ cuboids. When ExpressionFlow is equipped with ℓ_2 -norm, it is not difficult to see that for each cuboid, the computation complexity is $O(\lfloor \frac{X}{X} \rfloor \lfloor \frac{Y}{Y} \rfloor \lfloor \frac{T}{T} \rfloor)$. Therefore, the complexity of the whole algorithm is $O(XYT)$. In other words, it is linear to the size of the video sequence.

For ExpressionFlow equipped with ℓ_1 -norm, it runs Nelder-Mead simplex algorithm in $X \times Y \times T$ cuboids. However, little is known of the convergence of the optimization algorithm [18].

As for DTSA, a microexpression sequence is first reshape to $X \times X \times X$. And then the i -th mode tensor multiplication is performed on three modes in order to project it to a $L \times L \times L$ tensor. The process includes **unfolding-matrix multiplication-folding** process, leading to the complexity of $O(X^3L)$.

As for LBP-TOP, a microexpression sequence is first split into $X \times Y \times T$ blocks. Within each block, LBP-TOP histogram is extracted, which costs $O(\lfloor \frac{X}{X} \rfloor \lfloor \frac{Y}{Y} \rfloor \lfloor \frac{T}{T} \rfloor N)$, where N (typically set to 8) is the neighborhood parameter in the approach. This gives the overall complexity of $O(XYTN)$.

As for FDM, the time consumption comprises two parts, i.e. the optical flow estimation and the iterative principal direction computation. The latter was able to finish in several rounds. The main workload comes from the former, which is a numerical optimization process.

We will show the actual runtime of those methods in Sect. 8.4.4.

8.4 Experiments

In this section, we briefly describe several benchmark microexpression datasets and the experimental setup. Then a comprehensive comparison is performed between the proposed ExpressionFlow and other state-of-the-art methods in microexpression recognition. Finally, we discuss the effect of parameters and report the runtime of each method.

8.4.1 Experiments Setup

There are several microexpression datasets used in existing research works, including SMIC [21], SMIC2 [19], CASME I [40] and CASME II [41]. Among them, SMIC2 consists of three sub-datasets, i.e., HS, VIS, and NIR. Their abbreviations indicate the environments in which the datasets were captured, standing for the high-speed camera, normal visual camera, and near-infrared camera, respectively.

Figure 8.2 illustrates the data samples of each sub-dataset. Naturally, it is not easy to perceive the facial action change by bare eyes.



Fig. 8.2 Samples of cropped faces from CASME I [40] and CASME II [41]. The upper row shows an example of disgust from CASME I [s8-EP12-11-1]. The original sample contains 17 frames. Every second image is shown. The lower row shows an example of happiness from CASME II [s12-EP03-04]. The original sample contains 97 frames. Every ninth image is shown

To evaluate the performance of different approaches, the leave-one-out cross-validation scheme is applied. This method involves, for a given dataset, excluding all clips from one individual for testing purposes, while utilizing the remainder for training. This configuration aligns with practical applications, wherein the subject whose emotions we aim to analyze is not included in the training dataset.

The comparison is evaluated by using both accuracy and macro-mean $F1_M$ score [31] as the recognition performance. In particular

$$\text{Accuracy} = \frac{\sum_i (TP_i + TN_i)}{\sum_i (TP_i + TN_i + FP_i + FN_i)}, \quad (8.14)$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively, and

$$F1_M = \frac{2 \cdot \text{Precision}_M \cdot \text{Recall}_M}{\text{Precision}_M + \text{Recall}_M}, \quad (8.15)$$

where $\text{Precision}_M = \frac{1}{L} \sum_{i=1}^L \frac{TP_i}{TP_i + FP_i}$ and $\text{Recall}_M = \frac{1}{L} \sum_{i=1}^L \frac{TP_i}{FN_i + FP_i}$. L is the number of classes in a particular classification problem. Precision_M and Recall_M are the macro-mean precision and macro-mean recall, respectively. They treat all classes equally and thus avoid the problem of imbalanced data distribution.

There are two tasks in SMIC and SMIC2, i.e., identification and categorization. As for SMIC, the categorization task contains only positive and negative microexpressions. As in SMIC2, the identification task is identical to the one in SMIC. However, in the categorization task, it contains an additional surprise category.

CASME I and CASME II contain only the categorization task. Both tasks have more emotion classes than SMIC and SMIC2. We rule out the classes whose samples are insufficient (for example, there is only one contempt clip in CAMSE I, which would be impossible to predict if it is not in the training set). As a result, in our experiments, CASME I contains disgust, repression, surprise, and tenseness; CASME II contains disgust, happiness, repression, surprise, and others.

8.4.2 Recognition Results

We compare our methods against the FDM, two state-of-the-art methods, i.e., the method based on LBP-TOP [19], the method based on DTSA [32], as well as another advanced optical flow based approach, i.e., the Large Displacement Optical Flow (LDOF) [3].

In FDM, feature parameters are set to the same as in [38]. That is, the space division includes 4×4 , 8×8 , 16×16 , and 32×32 . And temporal multiplicity includes 2, 3 and 4.

In LBP-TOP, video clips are first interpolated to a fixed length of frames and then segmented into $X \times Y \times T$ blocks. Within each block, the descriptors are extracted and linked together to form the final representation. A SVM with an RBF kernel is used for microexpression classification. We use the parameter combination from the original paper, i.e., $5 \times 5 \times 1$, $5 \times 5 \times 2$, $8 \times 8 \times 1$, and $8 \times 8 \times 2$. The original implementation of LBP-TOP is used.

As in the DTSA-based method, face images are first rescaled to 64×64 pixels, and then each clip is temporally aligned to 64 frames. That is to say, any sample in the datasets is rescaled to a 3-order tensor $S \in \mathbb{R}^{64 \times 64 \times 64}$. Three 2-order tensors (matrices) $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3 \in \mathbb{R}^{64 \times K}$ ($K < 64$) are learned for projecting S to a low-dimensional space: $\tilde{S} = S \times_1 \mathcal{U}_1 \times_2 \mathcal{U}_2 \times_3 \mathcal{U}_3, S \in \mathbb{R}^{K \times K \times K}$. The learning process is carried out so that samples in the same category are near each other and samples from different categories are apart. A neural network algorithm, Extreme learning machine (ELM) [16], is employed for training the classification model. For a microexpression query, the clip is first rescaled to $\mathbb{R}^{64 \times 64 \times 64}$, and projected with the above-learned matrices. Classification processes are performed with the pre-learned neural network.

As for our method ExpressionFlow, several parameter combinations have been tested. The space divisions are selected among 4×4 , 6×6 , 8×8 , and 12×12 . The temporal divisions are selected among 2, 3, and 4.

We also experimented with an alternative optical flow-based method. The Large Displacement Optical Flow (LDOF) [3] encodes large movement information which is hardly captured by conventional methods. We reckon this approach may be able to differentiate between microexpression and non-microexpression since the movement magnitudes vary. We first apply the Temporal Interpolation Model to interpolate the sequence to 3, 5, or 7 frames, denoted TIM3, TIM5, and TIM7 in the following texts. The number of frames is selected as small as possible so that the difference between consecutive frames can be maximized, making LDOF better extract informative features.

Figure 8.3 shows the detailed $F1_M$ score and accuracy of the six candidates. Several comparisons can be made from the table:

- *EF vs. FDM.* It is obvious that both of the two ExpressionFlow methods exceed FDM in prediction performance, indicating that our improvement is effective. We have run a t -test [2] on the accuracies of both EF variants and FDM, and both variants are significantly better than FDM with a p-value of 0.0047 and 0.0035.

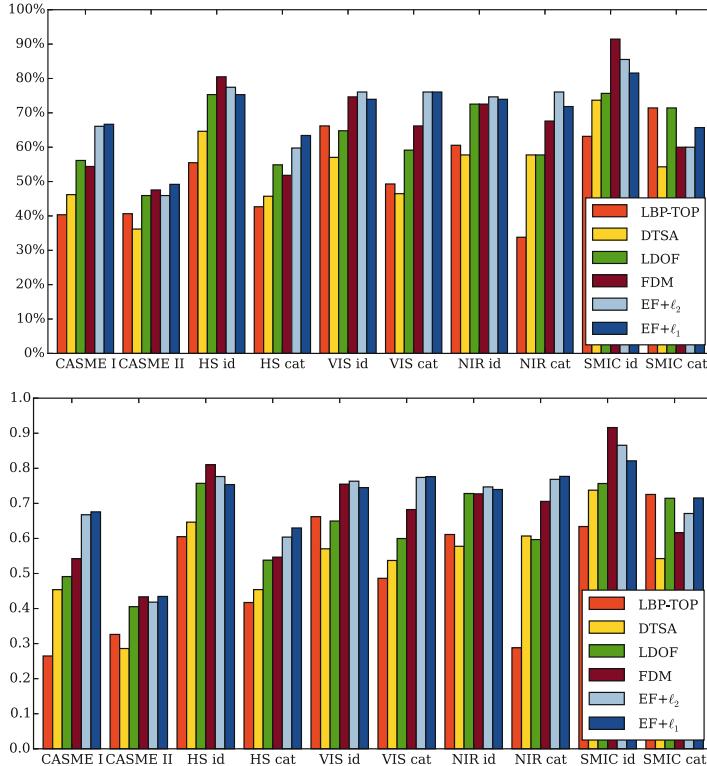


Fig. 8.3 Best results obtained with different approaches. Top: best results measured in accuracy; Bottom: best results measured in $F1_M$. Abbreviations: ‘ide’ for identification, ‘cat’ for categorization

- *EF vs. LBP-TOP, DTSA.* Also, the ExpressionFlow-based methods outperform LBP-TOP and DTSA. Both variants are significantly better than LBP-TOP and DTSA with the following p-values: 0.0013 (EF+ ℓ_2 vs. LBP-TOP), 1.49×10^{-7} (EF+ ℓ_1 vs. DTSA), 0.0011 (EF+ ℓ_1 vs. LBP-TOP), 6.30×10^{-7} (EF+ ℓ_1 vs. DTSA).
- *EF+ ℓ_1 vs. EF+ ℓ_2 .* On average, the ExpressionFlow equipped with ℓ_1 -norm target function attains slightly better recognition performance than the ExpressionFlow with ℓ_2 -norm target function. A possible reason is that the latter one is more sensitive to outliers, as claimed in Sect. 8.3.2.
- *EF vs. LDOF.* A straightforward application of LDOF achieves quite good results. However, most of the good performances of LDOF are on identification tasks. This is because an important difference between microexpression and non-microexpression is the movement magnitude. Since LDOF is designed to capture large displacement, it is not queer for LDOF to have such a good performance. Nevertheless, the advantage (only on identification tasks) is not significant under



Fig. 8.4 Samples of extracted ExpressionFlow (best read in color). Row 1: image sequence of the surprise microexpression from the HS dataset (hs-s3-sur-05). The sequence contains 28 images. Only 3 of them are shown for clarity. Row 2: the ExpressionFlow extracted with segmentation scheme [8, 8, 2]. Each grid corresponds to a cuboid in the original video. The hue of each grid indicates the direction of the motion vector in the corresponding cuboid, which can be looked up from the color wheel. The intensity of each grid indicates the magnitude of motion

a *t*-test. Meanwhile, EF-based methods are still significantly better than LDOF with a p-value of 0.0108 ($EF+\ell_2$) and 0.0033 ($EF+\ell_1$) on categorization tasks.

Figure 8.4 shows an example of extracted ExpressionFlow. In the example, the image sequence contains a surprise microexpression from the HS dataset. The ExpressionFlow is extracted with segmentation scheme [8, 8, 2]. Thus there are two ExpressionFlow frames, each of which is of size 8×8 . The hue of each grid indicates the direction of the motion vector in the corresponding cuboid, which can be looked up from the color wheel. The intensity indicates the magnitude of motion. As can be seen from the first frame, the orange cluster at the top left indicates the cuboids move towards the upper right. The green cluster at the top right indicates the cuboids move towards the upper left. The second frame can be interpreted in the same way. Read from the ExpressionFlow, the subject frowned his eyebrows and released, which is in accordance with our observation. In addition, as indicated by the second ExpressionFlow frame, the subject also showed a slight upward movement in the nose and mouth area.

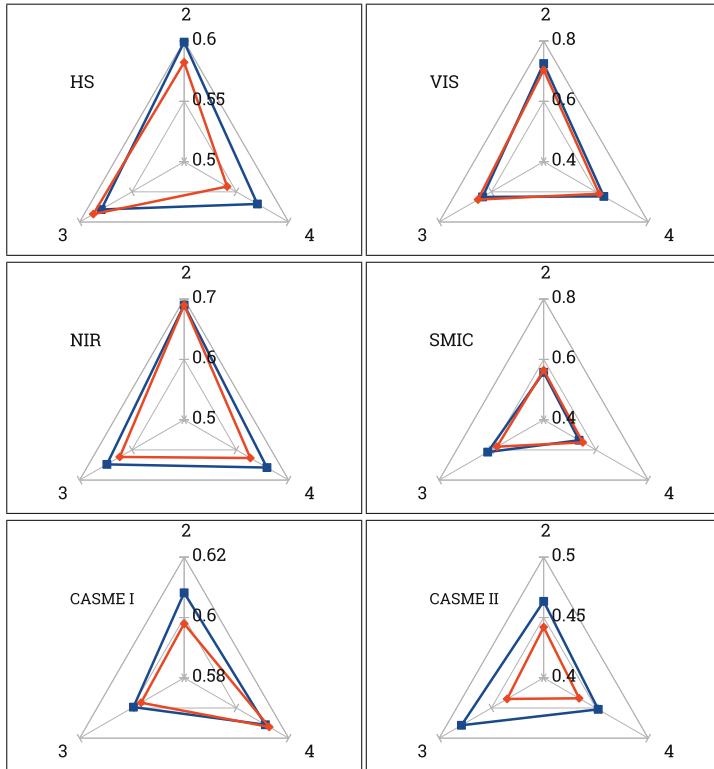


Fig. 8.5 Effect of temporal division on recognition results. Each spoke of a chart stands for accuracy obtained under a choice of T ($T \in \{2, 3, 4\}$). Blue lines: ExpressionFlow with ℓ_1 -norm; Red lines: ExpressionFlow with ℓ_2 -norm. Taking the upper-left chart for example, the blue dot on the ‘2’-spoke indicates the best accuracy of EF+ ℓ_1 reaches 0.6 when $T = 2$. While the accuracy of EF+ ℓ_2 in red dot when $T = 2$ is below 0.6 but above 0.55

8.4.3 Effects of Parameters

Figure 8.5 is a radar chart showing the best accuracies obtained from the different choices of t . From the figure, it can be seen that the performance degrades remarkably when $T = 3, 4$ in VIS and NIR. In other scenarios, meanwhile, the performance fluctuates normally. To explain this phenomenon, the frame rate of each dataset is studied. VIS and NIR have the lowest frame rate, i.e., 25 fps. It means that for a typical microexpression sequence, it lasts for 0.3 seconds and thus VIS and NIR capture 7–8 frames. When the temporal division is set to 3 or 4, each batch will only have 1–2 frames. This causes our target function ‘less determined’, and more likely to be influenced by noise or outliers. While other datasets have higher frame-rate, for example, the 100-fps HS dataset. It can have up to 7–10 frames in each batch, providing enough robust conditions to extract an accurate ExpressionFlow.

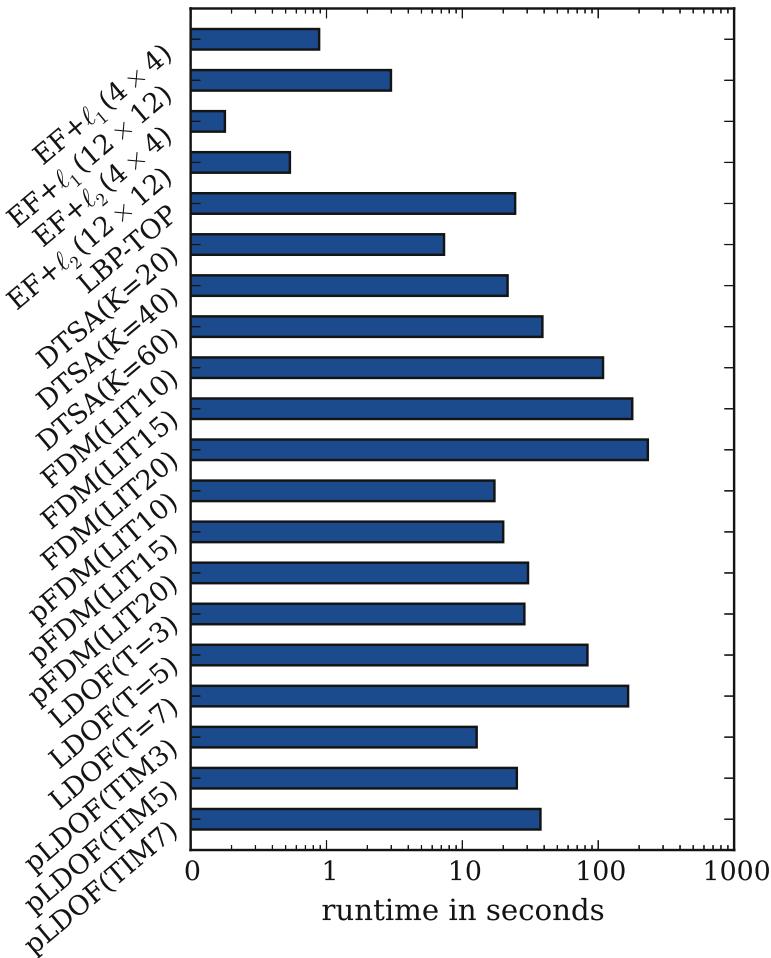


Fig. 8.6 Typical runtime of different methods processing a microexpression sequence in seconds. Extensive experiments have been conducted, but only parameters with a significant impact on runtime are shown in parentheses. Two variants of ExpressionFlow: space division; DTSA: reduced dimension; FDM and parallel FDM (pFDM): number of frames interpolated by the linear model; LDOF and parallel LDOF (pLDOF): number of frames interpolated by Temporal Interpolation Model. Please note the horizontal axis is in log scale

8.4.4 Runtime

Figure 8.6 compares the runtime of different methods processing a microexpression sequence.

ExpressionFlow with ℓ_1 -norm performs slightly better than its ℓ_2 -norm counterpart. The runtime of ℓ_2 -norm is much less than ℓ_1 -norm since no iterations are

involved. Therefore, it is recommended to use ExpressionFlow with ℓ_2 -norm in real-world applications.

As for DTSA, we show the amortized runtime for a single sequence. The runtime is sensitive to the dimension of the target tensor (the K parameter).

Both FDM and LDOF are sensitive to the number of frames in the sequence. In both approaches, parallel execution can largely reduce the time consumption.

ExpressionFlow with ℓ_2 -norm is faster than other competitors here. On average, EF+ ℓ_2 can process a microexpression in about 0.3 seconds. That is to say, we can recognize microexpressions in real time.

8.5 Conclusions and Future Work

We introduced an efficient descriptor, the ExpressionFlow, for microexpression analysis. Specifically, we split a microexpression sequence into multiple small cuboids, with each a motion vector calculated by minimizing our proposed objective function. The descriptor demonstrates a state-of-the-art recognition rate and can be processed in real time. Moreover, our approach captures the motion pattern of facial muscle actions, providing a true reflection of the essence of microexpressions.

In the future, we plan to optimize the segmentation algorithm for fully automatic microexpression recognition. Furthermore, a seamless system that combines microexpression and macroexpression recognition can be attractive. For instance, AU detection (Sect. 7.1) is an important sub-task in macroexpression recognition, but the detection of AUs in microexpressions has struggled. However, AU detection is still valuable in microexpression recognition. The correct AU recognition can serve as powerful supporting evidence for the emotion recognition results, thus enhancing interpretability. In addition, various properties of microexpressions are still being investigated in psychology, with many findings still being disputed. Fine-graded AU recognition, combined with visual calibration technology, could provide an important foundation for microexpression research in psychology and significantly contribute to cross-disciplinary auxiliary research.

References

1. Bernstein, D.M., Loftus, E.F.: How to tell if a particular memory is true or false. *Perspect. Psychol. Sci.* **4**(4), 370–374 (2009)
2. Box, J.F.: Guinness, Gosset, Fisher, and small samples. *Stat. Sci.* **2**, 45–52 (1987)
3. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(3), 500–513 (2011)
4. Calvo, R.A., D'Mello, S.: Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* **1**(1), 18–37 (2010)
5. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. *Comput. Vis. Image Underst.* **61**(1), 38–59 (1995)

6. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(6), 681–685 (2001)
7. Dacre Pool, L., Qualter, P.: Improving emotional intelligence and emotional self-efficacy through a teaching intervention for university students. *Learn. Individ. Differ.* **22**(3), 306–312 (2012)
8. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. *Psychiatry* **32**(1), 88–106 (1969)
9. Ekman, P., Friesen, W.: Facial Action Coding System. Consulting Psychologists Press, Palo Alto (1977)
10. Frank, M.G., Ekman, P.: The ability to detect deceit generalizes across different types of high-stake lies. *J. Pers. Soc. Psychol.* **72**(6), 1429 (1997)
11. Frank, M., Herbasz, M., Sinuk, K., Keller, A., Nolan, C.: I see how you feel: training laypeople and professionals to recognize fleeting emotions. In: Proceedings of the Annual Meeting of the International Communication Association (2009)
12. Gottman, J.M., Levenson, R.W.: A two-factor model for predicting when a couple will divorce: exploratory analyses using 14-year longitudinal data. *Fam. Process* **41**(1), 83–96 (2002)
13. Gottschalk, L.A., Auerbach, A.H., Haggard, E.A., Isaacs, K.S.: Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In: *Methods of Research in Psychotherapy*, pp. 154–165. Appleton-Century-Crofts, New York (1966)
14. Guo, Y., Xue, C., Wang, Y., Yu, M.: Micro-expression recognition based on CBP-TOP feature with elm. *Optik-Int. J. Light Electron Opt.* **126**(23), 4446–4451 (2015)
15. Happy, S., Routray, A.: Automatic facial expression recognition using features of salient facial patches. *IEEE Trans. Affect. Comput.* **6**(1), 1–12 (2015)
16. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. *Neurocomputing* **70**(1), 489–501 (2006)
17. Huang, X., Zhao, G., Hong, X., Zheng, W., Pietikäinen, M.: Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. *Neurocomputing* **175**, 564–578 (2016)
18. Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E.: Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM J. Optim.* **9**(1), 112–147 (1998)
19. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikainen, M.: A spontaneous micro-expression database: inducement, collection and baseline. In: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (2013)
20. Matsumoto, D., Hwang, H.S.: Evidence for training the ability to read microexpressions of emotion. *Motiv. Emot.* **35**(2), 181–191 (2011)
21. Pfister, T., Li, X., Zhao, G., Pietikainen, M.: Recognising spontaneous facial micro-expressions. In: Proceedings of the IEEE International Conference on Computer Vision (2011)
22. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expression detection in hi-speed video based on facial action coding system (FACS). *IEICE Trans. Inf. Syst.* **96**(1), 81–92 (2013)
23. Porter, S., Brinke, L.t.: Reading between the lies: identifying concealed and falsified emotions in universal facial expressions. *Psychol. Sci.* **19**(5), 508–514 (2008)
24. Russell, T.A., Chu, E., Phillips, M.L.: A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool. *Br. J. Clin. Psychol.* **45**(4), 579–583 (2006)
25. Salter, F., Grammer, K., Rikowski, A.: Sex differences in negotiating with powerful males. *Hum. Nat.* **16**(3), 306–321 (2005)
26. Salter, F., Grammer, K., Rikowski, A.: Sex differences in negotiating with powerful males. *Hum. Nat.* **16**(3), 306–321 (2005)
27. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: Proceedings of the International Conference on Computer Vision (2009)
28. Shreve, M., Godavarthy, S., Manohar, V., Goldgof, D., Sarkar, S.: Towards macro- and micro-expression spotting in video using strain patterns. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (2009)

29. Shreve, M., Godavarthy, S., Goldgof, D., Sarkar, S.: Macro-and micro-expression spotting in long videos using spatio-temporal strain. In: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (2011)
30. Tassinary, L.G., Cacioppo, J.T.: Unobservable facial actions and emotion. *Psychol. Sci.* **3**(1), 28–33 (1992)
31. Van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworth, London (1979)
32. Wang, S.J., Chen, H.L., Yan, W.J., Chen, Y.H., Fu, X.: Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. *Neural Process. Lett.* **39**(1), 25–43 (2014)
33. Wang, S.J., Yan, W.J., Li, X., Zhao, G., Fu, X.: Micro-expression recognition using dynamic textures on tensor independent color space. In: Proceedings of the International Conference on Pattern Recognition (2014)
34. Wang, S.J., Yan, W.J., Zhao, G., Fu, X., Zhou, C.G.: Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In: Proceedings of the European Conference on Computer Vision Workshops (2014)
35. Warren, G., Schertler, E., Bull, P.: Detecting deception from emotional and unemotional cues. *J. Nonverbal Behav.* **33**, 59–69 (2009)
36. Whitehill, J., Serpell, Z., Lin, Y.C., Foster, A., Movellan, J.R.: The faces of engagement: automatic recognition of student engagement from facial expressions. *IEEE Trans. Affect. Comput.* **5**(1), 86–98 (2014)
37. Wu, Q., Shen, X., Fu, X.: The machine knows what you are hiding: an automatic micro-expression recognition system. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction (2011)
38. Xu, F., Zhang, J., Wang, J.Z.: Microexpression identification and categorization using a facial dynamics map. *IEEE Trans. Affect. Comput.* **8**(2), 254–267 (2017)
39. Yan, W.J., Wu, Q., Liang, J., Chen, Y.H., Fu, X.: How fast are the leaked facial expressions: the duration of micro-expressions. *J. Nonverbal Behav.* **37**, 217–230 (2013)
40. Yan, W.J., Wu, Q., Liu, Y.J., Wang, S.J., Fu, X.: CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces. In: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (2013)
41. Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X.: CASME II: an improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* **9**(1), e86041 (2014)
42. Yan, W.J., Wang, S.J., Liu, Y.J., Wu, Q., Fu, X.: For micro-expression recognition: database and suggestions. *Neurocomputing* **136**, 82–87 (2014)
43. Zhang, P., Ben, X., Yan, R., Wu, C., Guo, C.: Micro-expression recognition system. *Optik-Int. J. Light Electron Opt.* **127**(3), 1395–1400 (2016)
44. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 915–928 (2007)
45. Zheng, W.: Multi-view facial expression recognition based on group sparse reduced-rank regression. *IEEE Trans. Affect. Comput.* **5**(1), 71–85 (2014)
46. Zhou, Z., Zhao, G., Pietikainen, M.: Towards a practical lipreading system. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2011)

Chapter 9

Emotion in the Neutral Face: Applications for Computer Vision and Aesthetics



Daniel N. Albohn and Joseph C. Brandenburg

Abstract Oftentimes “neutral” is classified as a baseline for other emotive categories such as angry, happy, or sad. Yet, neutrality has been the least studied amongst these discrete emotions. In this chapter, we focus on how neutrality is extracted, represented, and influences face perception and draw inferences to emotion perception in general. Neutral faces, or at least non-expressive faces, appear to be the most ubiquitous faces one encounters. Rarely does one experience a prototypical emotional face outside of a sterile lab or posed environment. This chapter explores the perceptual uniqueness of a neutral visage and the importance that understanding such an ambiguous face has across fields such as psychology and computer vision. We trace the psychological and computer science evolution of appreciating how a neutral face can inform and enhance the classification, detection, and judgment of faces that vary in characteristics such as age, race, and gender as well as incidental cues such as emotion, health, and attractiveness. This chapter summarizes classical and contemporary computer vision techniques as applied to (neutral) faces and discusses how a deeper understanding of non-emotionality is paramount to furthering the study of aesthetics, emotionality, and artistic value of people, objects, and scenes across multiple domains.

9.1 Introduction

The popular idiom, “don’t judge a book by its cover” simply means that despite what can readily be seen, the contents may not reflect the external appearance of the book and deserves appreciation on its own. This particular vernacular

D. N. Albohn (✉)

Booth School of Business, The University of Chicago, Chicago, IL, USA
e-mail: daniel.albohn@chicagobooth.edu

J. C. Brandenburg

Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, University Park, PA, USA
e-mail: jcb5590@psu.edu

expression has been used, for example, by countless parents to stymie any negative impressions a child forms of another individual without due process. The sentiment behind this statement also echoes through myriad psychological phenomena studied within person perception. For instance, individuals appear able to deduce sexual orientation, political affiliation, and some mental health disorders from face images alone [14, 34, 46]. In a particularly striking example of judging a book by its cover, one study examined whether individuals were able to determine stable emotion disposition (i.e., what the individual is actually feeling) from neutral face images of older adults [2, 27]. Incredibly, the individuals in these studies were able to accurately predict how positive or negative older adults (women in particular) were feeling simply from a static image of their neutral face. In this instance the individuals were quite literally judging the book (an individual) by its cover (their face).

By stating, “don’t judge” the expression implies that one is capable of not judging the book’s cover. This is likely an impossible feat. As a lighthearted example, consider a book with a completely blank, white cover. Arguably, there is nothing to judge the contents of the book by and therefore one might think that a judgment cannot be passed. However, the mere fact that the book cover is blank invites a whole host of questions that will influence perception and ultimately judgment. “Are the contents of the book so good that there is no need for a cover?” or, “The blank cover means that the protagonist experiences an existential crisis!”. This tongue-in-cheek example begs a larger question: Is a blank cover really “neutral”? Does “blankness”, “neutrality”, or “emptiness” allow for no judgment at all to take place? Indeed, if any judgment occurs the book cover is ostensibly not “neutral”! That is, a true “neutral” book cover would render the perceiver without any descriptive prose.

In this way, impressions derived from visual perception are a bit of a catch-22. In order to pass a judgment of, say, “neutral”, one has to first view a stimulus, but as soon as it is perceived its properties will necessarily influence judgment away from “neutral.” Sometimes the stimulus has properties that are easily binned into categories that are associated with specific judgments, such as a face with an anger expression being perceived as “negative” and “threatening.” On the other hand, sometimes the stimulus has properties that are less easily associated with specific judgments, such as a wooden chair. However, even in such an ambiguous case, individuals appear able to effortlessly draw inferences from such objects—perhaps the color or texture of wood is evocative of a country aesthetic which the perceiver associates with positivity. In both examples, the perceiver rendered an affective judgment despite one stimulus being relatively emotionally unambiguous and the other being largely emotionally ambiguous. The irony of “neutrality” is that humans despise ambiguity so much that they will go out of their way to find meaning where there is little to none, even if that means fabricating it all together. The psychologist Thomas Gilovich succinctly sums this up by stating: “[Humans] are predisposed to see order, pattern, and meaning in the world, and we find randomness, chaos, and meaninglessness unsatisfying. Human nature abhors a lack of predictability and the absence of meaning. As a consequence, we tend to “see” order where there

is none, and we spot meaningful patterns where only the vagaries of chance are operating” [20, p. 9].

Gilovich was reflecting on broader fallacies present in everyday life, but his sentiment applies more narrowly to the ambiguity of neutrality as well. In fact, we argue in this chapter that the inherent ambiguity of neutrality—whether it be in faces, art, fashion, etc.—is what makes it so unique. Ambiguous “neutral” stimuli allow for the individual to easily project their own biases into their judgments, which in turn can reveal interesting patterns of behavior not observable when a stimulus already comes “pre-loaded” with strong affective content (e.g., faces with emotional expressions, a beautiful landscape photograph, or a bowl of maggots). In other words, examining ambiguous “neutral” stimuli allows researchers to trace behavior as it unfolds from the properties of the stimulus itself through downstream perceiver outcomes unconstrained from any normative expectations about the stimulus.

While we focus this chapter on “neutral” within the framework of person perception, the central thesis of our argument is that insights and methods from this research can be applied and yield interesting insights into perception more broadly, including perceptual aesthetics across many domains. Similarly, we utilize the term “person perception” loosely to denote judgments provided by people rather than algorithms. In this way, we incorporate research not only from the emotion and face perception literature, but also from object and scene perception as well. When we discuss neutral perception by machines or algorithms, we explicitly label this as “computer vision” or “machine-derived judgments”.

With these caveats in mind, the remainder of this chapter will highlight some important contributions the study of “neutral” has provided to person perception and computer vision, with a focus on affective judgments derived from (non-)neutral stimuli. We first review literature attempting to operationalize the term “neutral”. Next, we discuss how “neutral” has been used in person perception to inform judgments and the applications that this has for computer vision. Finally, we close with how we believe this work can be informative for the field of general aesthetics and offer some simple suggestions to consider for future work. Throughout it all, we highlight the unique properties of “neutral” and “neutral stimuli” that set it apart from others.

9.2 Definitions of “Neutral”

Take a moment to think about whether you have ever encountered a truly neutral face and whether it provided any social information to help you make a judgment. When we asked participants to tell us what social information could be reliably derived from neutral faces, the majority of participants said “nothing,” while a minority provided explanations related to gender-emotion stereotypes or detailed backstories to support their inferences. One participant wrote, “A neutral face can sometimes give off an angry impression or like the girl is mad. This is pretty common among

young female emotionless faces”, and another wrote “She experiences so many challenges which have not taken hers [*sic*] down. She is tranquil all the time” [5].

These examples bring up two important considerations to keep in mind as we attempt to operationalize neutral. First, the majority of laypersons believe that neutral faces are uninformative and do not influence their judgments in any meaningful way. However, as mentioned earlier (and will show evidence for in the next section), people appear unable not to judge a face even when it is not overtly expressing any information. Second, the examples show that despite majority consensus about neutral faces, there is still considerable variation in how useful people view such faces when making judgments. Thus, what one individual classifies as a neutral visage may not be what another individual—or group of individuals—considers neutral. These two features of neutrality, which we believe not to be unique to faces, make studying it both incredibly difficult but also an incredibly useful tool for scientists. Despite these potential limitations, attempting to operationalize neutrality is important in order to fully appreciate its utility.

While it may seem rather obvious what constitutes “neutral” or “neutrality,” both laypersons and experts alike cannot agree on a common definition, rendering operationalization difficult. For instance, focusing solely within the domain of emotion and face perception, there are several definitions of neutral. For instance, Gasper et al. [18] contended that “neutral” is theoretically independent from “positive” or “negative” affective states and that it indicates an inattention or non-preference. Whereas Ekman and colleagues characterized neutral as the “baseline for the actor” [16, p. 73], alluding to the fact that “neutral” varies idiosyncratically. However, others place “neutral” emotion or faces at the midpoint within a two-dimensional space consisting of an arousal dimension (weak to strong) and a valence dimension (negative to positive) [33]. Russell [35] further clarified that the intersection of this two-dimensional arousal-valence space represented an “average everyday feeling” rather than pure “neutral” (p. 501), suggesting that the midpoint varies not only by the individual but also by what is being evaluated. Further corroborating Russell’s observation, Carrera-Levillain and Fernandez-Dols [12] found that perceptions of neutral expressions were found to be around the midpoint of arousal-valence space rather than exactly at the midpoint when compared to expressive faces. Likewise, Shah and Lewis [39] used multidimensional scaling to locate where perceptions of neutral were located in two-dimensional space relative to other expressions. Their model indicated that neutral was not in the middle, but rather at the periphery alongside other mildly negative emotions such as bored and tired [39]. We followed this line of thought and examined people’s mental representations (i.e., what they imagine in their “mind’s eye”) of neutral faces using a technique called reverse correlation. Our results showed that people’s mental representations of neutral faces indeed contained a small degree of (negative) expressivity, again suggesting that a truly “neutral” face is not devoid of emotionality [7]. Figure 9.1 shows an example of stimuli produced through this procedure. Interestingly, while the aggregate reverse correlation neutral face was rated more negative compared to foils, there was variability in what each individual conjured in their mind’s eyes as

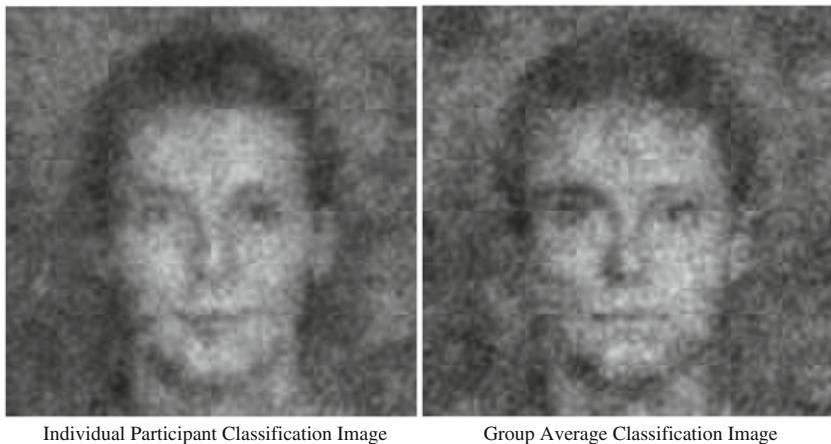


Fig. 9.1 Aggregate reverse correlation image of a typical “neutral” face (right) and an individual participant CI (left)

“prototypical neutral” as evidenced by clear smiliness in the example face on the left.

The lack of an empirically validated, agreed upon definition of neutral emotion and neutral faces makes it difficult to compare and contrast results across research programs. Is neutral the exact midpoint on a two-dimensional array? Or is it anything that closely orbits that midpoint? Is it even appropriate to place neutrality on such a plane in the first place? All considered, the lack of a common meaning of neutral, highlighted by multiple definitions within emotion and face perception, underscores the difficulty in studying “neutral” stimuli. However, across all of these examples, there appears to be two common operational parts to a formal definition of “neutral.” First, neutral appears idiosyncratic. What is “neutral” for one individual is not necessarily “neutral” for another person. Second, it appears that a prototypical neutral (face) does not exist. Instead, what people conceptualize as “neutral” actually contains a small amount of emotionality (though in line with the first point this likely varies by individual). In other words, what individuals report and presume to be “neutral” actually contains a low-level, but measurable, amount of valence.

Throughout the remainder of this chapter, we conceptualize neutrality based on these commonalities across multiple definitions. More formally, we suggest that perceptions of “true neutral” is theoretically impossible and responses to stimuli that are ostensibly neutral in appearance or contain ostensibly neutral properties are highly individualized.¹ In line with these assumptions is the phenomenon of

¹ While parts of this definition might apply to an individual’s experience or subjective display of neutral, the focus of this chapter is primarily on the perception of neutral. For a discussion about the experience of neutral as a feeling, we direct the reader to other sources (e.g., [19]).

micro-valances, the small—but measurable—emotional responses to objects that are commonly thought to be relatively mundane and emotionally “neutral” such as a tea pot or leather couch (see e.g., [25]).

In summary, while there may not be a consensus on an operational definition of neutral, the commonalities that we highlighted above provide a good scaffolding for conceptualizing the phenomenon, characteristics, and consequences of neutral across faces and other objects.

9.3 “Neutral” in Person Perception

In this section, we briefly discuss two important facets of perceiving neutral faces. Specifically, we focus on research that has examined both the “weirdness” and impossibility of true neutral. Both of these phenomena underscore neutral’s unique perceptual properties. Again, we focus on perception of neutral faces, but also draw on evidence from the person perception literature more generally (e.g., people judging objects, scenes).

9.3.1 *Neutral is Disturbing*

If a neutral face truly existed, perceivers would likely find it highly disturbing. Our facial muscles are always contracted to some degree and create a constant tension which holds the face in a specific configuration. Removing this muscle tension would result in an odd appearance, akin to the drooping of facial muscles and skin seen in individuals with facial paralysis. It has been shown that facial paralysis biases perceivers’ attention and disrupts normal face processing [15, 24]. Similarly, faces that appear “vacuous” and “empty” of any thought or content are less memorable compared to faces that demonstrate more emotional complexity [17].

The eeriness of expressionless, neutral faces is found in infants and adults alike. The still face paradigm is regularly used in research to induce stress in infants [44]. In this paradigm, adult caregivers suddenly switch from emoting to a blank, neutral face as they are interacting with their infant. When this change occurs, most infants become distressed and signal disapproval to their caregiver. Similarly, adults are disturbed when other entities such as robots and dolls possess human-like characteristics that lack the nuanced and naturally occurring human expressivity to match. This phenomenon is so prevalent that it has gained the name “The Uncanny Valley” [30], popularized through common television shows and movies such as *Westworld* and *I, Robot*. When expressionless faces are given emotional expressions, individuals are less disturbed by them, presumably because it shifts perception of the entity toward appearing more human-like compared to entities without expressions [10].

9.3.2 *True Neutral Is Impossible*

Setting aside the physiological impossibility of a “neutral face,” there are still other cues, features, and stereotypes that prevent neutral from existing. If there was a face that was devoid of emotionality it would also have to lack any other social cues such as gender, race, face shape, and context in order for individuals to not judge the “book by its cover.” In support of this, when individuals are asked to pose their best “neutral” expression, naive individuals’ perception of the subjectively-posed “neutral” expressions still appear to be influenced by factors such as stereotypes (e.g., emotion-resembling features, gender, race, and age), actual residual emotion, context, and other phenotypic face traits (e.g., brow to height width ratio, etc.).

Stereotypes about emotion play an important role in determining how a neutral face is evaluated. For instance, neutral faces on women compared to men are not evaluated equally. Men’s neutral faces are typically judged as angrier than those of women, whereas women’s neutral faces are typically judged as more fearful and joyful compared to neutral faces of men [1]. Further, androgynous faces expressing anger are more likely to be categorized as “male” whereas androgynous faces expressing fear and joy are more likely to be categorized as “female” [22]. Such differences in judgments are likely driven by naturally occurring phenotypic properties of sexually dimorphic faces (e.g., women’s eyes naturally being rounder and larger than men’s, mimicking cues common to fear expressions). Indeed, men are expected to show dominance and be overall more stoic. Thus, when men display a neutral face they are seen as more dominant than when women display neutral visages [21].

Stereotypes about age and race also alter perceptions of neutral faces. Older adult faces are judged by others—usually younger—as appearing more negative (angrier, sadder) and less positive (happy) than comparable younger adult neutral faces [5]. One explanation for this difference is that the aging properties of the face (e.g., wrinkles, sagging skin) mimic negative expressivity cues which perceivers mistake as actual emotional disposition [2, 23]. Similarly, Black neutral faces are typically judged as angrier and more threatening than White neutral faces [3]. However, this pattern of results is likely due to race stereotypes rather than the properties of the face itself as objective measures of Black and White faces show the opposite pattern—Black faces structurally resemble fear expressions more than White faces while White faces structurally resemble anger expressions more than Black faces [3, 48]. We discuss this counter-stereotypic effect further in the computer vision section below.

While stereotypes related to social identity and emotion typically result from emotion-resemblance or emotion stereotypes, there has also been work examining how actual emotion cues in neutral faces influence perception. In a series of studies, Albohn and Adams [4] found that subtle emotion cues remain on faces when a participant returns to a neutral expression. This “emotion residue” was not only detectable by others, but also influenced impression judgments similar to that of overt expressions. neutral faces that contained anger emotion residue (i.e., a neutral

face occurring after an individual made an anger expression and returned to neutral) were judged higher on negative attributes such as troublesome, rude, and uncivil. On the other hand, neutral faces containing happy emotion residue were judged higher on positive attributes such as smart, enthusiastic, and trustworthy. The existence of residual emotional tone on neutral faces that others are able to correctly identify and respond to underscores the notion that our faces may never truly be devoid of affective value and is never “neutral” in appearance.

Context in which the neutral stimulus appears also influences judgments of it. This effect has been well-documented across both faces, scenes, and objects. In an early examination on the relativity of facial expressions, Russell and Fehr [36] showed that judgements of neutral faces depended on the expressions that preceded it. For example, neutral faces were rated as happier, calmer, and more content if participants were first shown stimuli with sad and disgusted expressions. Likewise, when participants were first shown stimuli with calm and excited expressions, the neutral faces were rated as angrier, sadder, and more disgusted [36]. This work suggests that prior expressions act as a perceptual anchor that shifts judgements of neutral faces relative to the affective content of the previously seen expressions.

Others have found that neutral stimuli are more similar to negative expressions than positive ones. Lee et al. [26] reported that classification of neutral faces were more similar to negative (fear) expressions than positive (happy) expressions. Similarly, Tae et al. [42] found that neutral stimuli (both faces and scenes) were harder to distinguish from negative stimuli compared to positive stimuli.

Finally, a number of studies have attempted to parse the physical properties of neutral faces and objects that give rise to non-neutral judgments. Neth and Martinez [31] showed that the physical placement and configuration of face features such as the mouth and eyes could alter perception of neutral faces. They manipulated the physical distance between eyes and mouth on neutral faces. When the distance was increased participants categorized the face as sadder, whereas when the distance was decreased, participants categorized the face as angrier [31]. Similarly, Mignault & Chaudhuri [28] showed that when a neutral face was on a bowed, lowered head it was evaluated as more submissive and sad. On the other hand, neutral faces that are on raised heads were evaluated as displaying more dominant emotions.

9.4 “Neutral” in Computer Vision

Scientific research has been employing computer algorithms in order to compute and predict real life outcomes for decades. Advances within this field began to rapidly accelerate in the 1950s when computers were successfully trained to recognize images (perceptron), predict numbers (SNARC), and even play checkers (IBM). By the 1960s, theoretical advances started to outpace the limited computational power of technology at the time, limiting the tangible advances of the field until computer technology could catch up. Computer hardware at the time was incapable of handling the complexity and intensity of the computations required

for ingesting real life data [29]. Though these limits were in place, the theoretical underpinnings that laid the foundation for computer vision and machine learning continued to rapidly accelerate. By the start of the twenty-first century, computer vision and machine learning had already begun to penetrate commercial, research, and personal use. For example, by 1997 the U.S. Postal Service had widely adopted automated address recognition and letter sorting of handwritten parcels throughout their facilities, powered by machine learning tools trained to recognize digits and letters [38].

Computer vision has been used extensively to understand the complexities of the face. Early work focused on automated recognition and classification of lower-order expressive units of emotion expressions. The most well-researched type of expressive units are those belonging to the Facial Action Coding System (FACS) [16]. In FACS facial expressions are broken down into smaller facial movements (action units; AUs) that when combined create more complex states such as emotion expressions. Each of these facial action units roughly corresponds to an underlying facial muscle. For example, facial action unit 18 is called the “lip puckerer” which is activated when an individual puckers their lips (facial muscles Incisivii labii superioris and inferioris).

Facial AUs are well studied within the affective computing community as they are naturally well suited for ingestion by algorithms. Part of their appeal is that their ground truth relies less on the subjective rating by individuals and more on the objective measurement of physical properties of the face. Early efforts to automate the recognition of AUs utilized a variety of methods including optical flow, spatial, and physical measurement [9, 13]. Utilizing these methods, researchers were able to achieve facial AU recognition accuracy of around 90%, which was on par with accuracy rates of professional FACS coders. Contemporary models that combine many approaches or use deep learning can achieve AU recognition accuracies of over well over 90% (see [8]).

Beyond AUs, machine learning has been used to understand other properties of the face that might inform judgments. Previous work in behavioral science has shown that properties such as facial structure, color, and texture all exert an influence on perceptual judgments derived from faces (see e.g., [40]). These properties have also been studied using computer vision, albeit separately. For instance, when neural network classifiers are trained on overtly expressive faces, facial structural resemblance to anger expressions was correlated with threatening personality traits (e.g., dominant), and resemblance to happy expressions was correlated with positive traits (e.g., caring [37]). Yip and Sinha [47] showed that when shape cues are visually degraded by blurring an image, color cues become increasingly important to accurately recognize and categorize a face. Indeed, when participants were presented blurred gray-scaled images (degraded in both shape and color), participants’ accuracy dropped significantly, again underscoring the importance of color cues. David Perrett and colleagues have shown that an individual’s facial redness (via, e.g., the body’s carotenoids, which produces a yellow-orange skin coloration) has a direct influence on a number of human impressions of the individual. For example, facial redness has been shown to increase both men’s and

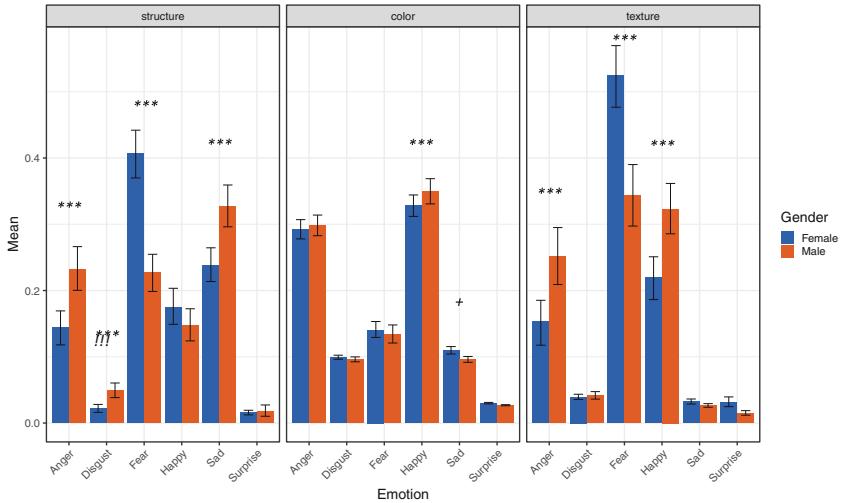


Fig. 9.2 Structure, color, and texture (panels) resemblance to each emotion by gender of neutral face

women's perceived attractiveness and health [32, 43]. Lastly, related to face texture, Tsankova and Kappas [45] manipulated the apparent "smoothness" of facial skin by digitally removing skin blemishes. Participants rated the altered, smooth-skin neutral faces as appearing more trustworthy, competent, attractive, and healthier.

More recently, Albohn and Adams [6] extended previous computer vision research by examining the separate and combined influence of structure, color, and texture similarity to emotion expressions in neutral faces. Albohn and Adams [6] trained three machine learning classifier models (one for each face metric) on several hundred emotional faces. Next, they applied the trained models to posed neutral expressions varying in gender. The output from these models revealed how similar in appearance men's and women's neutral faces were to overt expressions. As depicted in Fig. 9.2, men's neutral faces were typically more similar to dominant expression across structure, color, and texture (anger, happy) whereas women's neutral faces were more similar to submissive emotions (e.g., fear, sadness) across all three metrics. These results were in line with behavioral results and suggest that individuals use and integrate all three channels of information when forming judgments from neutral faces.

Similar computer vision models have been used to understand race-emotion stereotypes for Black and White neutral faces. In this work, the structural similarity of Black and White neutral faces to emotion expressions was estimated. Recall that in the previous section, we noted that individuals typically rate Black neutral faces as angrier than White neutral faces. However, results from computer vision algorithms demonstrated a counter-stereotypic pattern of results. White neutral faces were more similar to anger expressions whereas Black neutral faces were more similar to fear expressions [3, 48]. Interestingly, these results are opposite of what would be

expected given prior work on race-emotion stereotypes provided by individuals (i.e., not computer algorithms).

Taken together, the research on facial structure, color, and texture leverages computer vision to help disentangle the effects of top-down versus bottom-up influences on perception. In the case of gender-emotion stereotypes, it appears that the similarity a specific face property has to an overt expression (a bottom-up cue) and societal gender stereotypes (a top-down cue) are in alignment. That is, for example, men's neutral faces are objectively more similar to anger expressions and are judged as angrier compared to women's neutral faces. On the other hand, White neutral faces are objectively more similar to anger expressions compared to Black neutral faces, but stereotypes related to these social groups appear to override physical properties of the face resulting in a counter-stereotypic computer vision finding. This pattern of results suggest that top-down race-emotion stereotypes are so powerful that they appear to override the phenotypic similarity to emotion expressions and highlights the utility that more "objective" machine learning algorithms have for studying how individuals are perceived.

9.5 Bias in Computer Vision

Before we conclude, it is important to recognize the potential harm machine learning models can cause if utilized improperly or with less-than-ideal training data (see also Chapter 4 this volume). It has been well documented in the face detection literature that various algorithms have different performance accuracies depending on the gender and race of the face being evaluated [11]. The American Civil Liberties Union (ACLU) provided an eye-opening example of this bias when they conducted a study showing that commercially available face-matching recognition software (Amazon's "Rekognition") falsely matched 28 members of congress with actual criminal mugshots [41]. Perhaps more concerning, however, is that of the 28 false positives, 11 were people of color. While the results of the ACLU's investigations have already ignited a host of policy changes for the use of such technology by law enforcement officials, the underlying issue remains: bias can be trained into algorithms. This issue is of clear societal concern. Thus, it is important to examine race and gender effects of any algorithm and compare them with known human responses, given that such issues are pervasive in contemporary computer vision applications.

Whenever we utilize computer vision algorithms to inform research or interpret results that use machine learning, it is important to keep these limitations in mind. You might have noticed in the previous section we put objective in quotes when describing the machine learning models. Because most machine learning models are trained on human provided data, they necessarily inherit all of the noise and biases present in the individuals rating or categorizing the training data. The "objectivity" from these models is provided by isolating specific face features/metrics which allows determining their individual influence at the exclusion of others. In other

words, a model trained on only facial structure similarity to emotional expressions can only use that information to make predictions. On the other hand, if we asked a human participant to judge the structural similarity a neutral face had to a specific emotion it is likely impossible for that individual not to incorporate other properties to some degree. In this regard, the machine learning model provides a purer, more objective metric of “structural similarity” compared to human judgments.

9.6 Broader Implications

At this point, you may be asking yourself how the study of neutral faces can contribute to the field of perceptual aesthetics. Here we offer several suggestions on how evidence from behavioral and computational approaches to neutral face perception can inform the field of aesthetics. These suggestions are intentionally broad and are meant to offer a foundation for which to build more neutral-conscientious research.

First, we suggest that researchers redefine their definitions of “neutral.” Scientists should not assume that a purported “neutral” stimulus will be perceived as emotionally “in the middle” of whatever dimensional emotion space is being utilized. In this chapter, we reviewed evidence from both humans and machine algorithms that suggest that neutral faces are not evaluated as affectively neutral. Likewise, we also urge researchers to consider the idiosyncrasies that are apparent in judgments of neutral stimuli. While there are general trends in the perception of neutral (e.g., slightly negative) researchers should expect it to vary by individual. Sometimes these idiosyncrasies may be minuscule, but at other times they can alter judgments entirely. For example, see Fig. 9.2 where the individual participant neutral mental representation is clearly expressing positivity while the group average is expressing (and is rated) negativity.

Second, we suggest that researchers avoid using neutral stimuli as a baseline, midpoint, or control. Instead, neutral stimuli should be treated, analyzed, and evaluated as a separate stimulus condition, much like “positive” or “negative.” While neutral can be a perfectly suitable “control” condition in some experiments, researchers should keep in mind that affective responses to certain neutral stimuli may be more similar to one of their other conditions (e.g., negative) rather than perfectly in the middle.

Lastly, we want to underscore the interesting opportunities that studying neutral has for computer vision. While it appears possible to computationally predict how individuals will respond to neutral stimuli, such approaches should be treated as an examination of stimuli that can vary just as much as any other affective category.

9.7 Conclusions

Neutral is an abstract and difficult concept to precisely define within the perception literature. However, in this chapter we explored two commonalities across definitions that are important to keep in mind when studying the affective content of stimuli, whether it be through human observation or computer algorithms. First, we highlighted work from both behavioral science and computer vision that suggests the affective content of “neutral” faces, scenes, and objects are not evaluated as exactly at the midpoint between “negative” and “positive.” Second, while judgments of neutral stimuli tend to be closer to “negative” than “positive”, we also reviewed literature that suggests that there is some individual variability in how such stimuli are evaluated. Both of these aspects of neutral are important for researchers to keep in mind while studying or utilizing neutral stimuli in their work.

The inherent ambiguity of neutral expressions, scenes, and objects provide a unique opportunity for researchers to examine how individuals and machines respond to stimuli under obscure conditions. That neutral is ambiguous by nature, affords it the opportunity to “hijack” the visual system and influence impressions via other powerful predictors of impressions such as expression, emotion resemblance, gender, and context. Given this complexity, it is important to continue to evaluate and understand how computer vision and machine learning perform in relation to known cues that humans use to make judgements and form impressions.

References

1. Adams Jr., R.B., Nelson, A.J., Soto, J.A., Hess, U., Kleck, R.E.: Emotion in the neutral face: a mechanism for impression formation? *Cognit. Emot.* **26**(3), 431–441 (2012)
2. Adams Jr., R.B., Garrido, C.O., Albohn, D.N., Hess, U., Kleck, R.E.: What facial appearance reveals over time: when perceived expressions in neutral faces reveal stable emotion dispositions. *Front. Psychol.* **7**, 986 (2016)
3. Adams Jr., R.B., Albohn, D.N., Hedgecoth, N., Garrido, C.O., Adams, K.D.: Angry white faces: a contradiction of racial stereotypes and emotion-resembling appearance. *Affect. Sci.* **3**(1), 46–61 (2020)
4. Albohn, D.N., Adams Jr., R.B.: Emotion residue in neutral faces: implications for impression formation. *Soc. Psychol. Personal. Sci.* **12**(4), 479–486 (2020)
5. Albohn, D.N., Adams Jr., R.B.: Everyday beliefs about emotion perceptually derived from neutral facial appearance. *Front. Psychol.* **11**, 264 (2020)
6. Albohn, D.N., Adams Jr., R.B.: The expressive triad: structure, color, and texture similarity of emotion expressions predict impressions of neutral faces. *Front. Psychol.* **12**, 612923 (2021)
7. Albohn, D.N., Brandenburg, J.C., Adams Jr., R.B.: Perceiving emotion in the “neutral” face: a powerful mechanism of person perception. In: Hess, U., Hareli, S. (eds.) *The Social Nature of Emotion Expression*, pp. 25–47. Springer International Publishing, Berlin (2012)
8. Baltrušaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: OpenFace 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66. IEEE (2018)
9. Bartlett, M.S., Hager, J.C., Ekman, P., Sejnowski, T.J.: Measuring facial expressions by computer image analysis. *Psychophysiology* **36**(2), 253–263 (1999)

10. Bowling, N.C., Banissy, M.J.: Emotion expression modulates perception of animacy from faces. *J. Exp. Soc. Psychol.* **71**, 83–95 (2017)
11. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. *Proc. Mach. Learn. Res.* **81**, 1–15 (2018)
12. Carrera-Levillain, P., Fernandez-Dols, J.M.: Neutral faces in context: their emotional meaning and their function. *J. Nonverbal Behav.* **18**(4), 281–299 (1994)
13. Cohn, J.F., Zlochower, A.J., Lien, J., Kanade, T.: Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding. *Psychophysiology* **36**(1), 35–43 (1999)
14. Daros, A.R., Ruocco, A.C., Rule, N.O.: Identifying mental disorder from the faces of women with borderline personality disorder. *J. Nonverbal Behav.* **40**(4), 255–281 (2016)
15. Dey, J.K., Ishii, M., Boahene, K.D., Byrne, P.J., Ishii, L.E.: Facial reanimation surgery restores affect display. *Otol. Neurotol.* **35**(1), 182–187 (1994)
16. Ekman, P., Friesen, W.V.: Facial Action Coding System. American Psychological Association, Washington (1978). Type: dataset
17. Franklin, R.G., Adams Jr., R.B.: What makes a face memorable? the relationship between face memory and emotional state reasoning. *Pers. Individ. Differ.* **49**(1), 8–12 (2010)
18. Gasper, K., Hackenbracht, J.: Too busy to feel neutral: reducing cognitive resources attenuates neutral affective states. *Motiv. Emot.* **39**(3), 458–466 (2015)
19. Gasper, K., Spencer, L.A., Hu, D.: Does neutral affect exist? how challenging three beliefs about neutral affect can advance affective research. *Front. Psychol.* **10**, 2476 (2019)
20. Gilovich, T.: How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life, 1. Free Press Paperback edn. Free Press, New York (1993)
21. Hareli, S., Shomrat, N., Hess, U.: Emotional versus neutral expressions and perceptions of social dominance and submissiveness. *Emotion* **9**(3), 378–384 (2009)
22. Hess, U., Adams Jr., R.B., Grammer, K., Kleck, R.E.: Face gender and emotion expression: are angry women more like men? *J. Vis.* **9**(12), 19–19 (2009)
23. Hess, U., Adams Jr., R.B., Simard, A., Stevenson, M.T., Kleck, R.E.: Smiling and sad wrinkles: age-related changes in the face and the perception of emotions and intentions. *J. Exp. Soc. Psychol.* **48**(6), 1377–1380 (2012)
24. Ishii, L., Carey, J., Byrne, P., Zee, D.S., Ishii, M.: Measuring attentional bias to peripheral facial deformities. *Laryngoscope* **119**(3), 459–465 (2009)
25. Lebrecht, S., Bar, M., Barrett, L.F., Tarr, M.J.: Micro-valences: perceiving affective valence in everyday objects. *Front. Psychol.* **3**, 107 (2012)
26. Lee, E., Kang, J.I., Park, I.H., Kim, J.J., An, S.K.: Is a neutral face really evaluated as being emotionally neutral? *Psychiatry Res.* **157**(1), 77–85 (2008)
27. Malatesta, C.Z., Fiore, M.J., Messina, J.J.: Affect, personality, and facial expressive characteristics of older people. *Psychol. Aging* **2**(1), 6 (1987)
28. Mignault, A.: The many faces of a neutral face: head tilt and perception of dominance and emotion. *J. Nonverbal Behav.* **27**(2), 111–132 (2003)
29. Minsky, M., Papert, S.A.: Perceptrons: An Introduction to Computational Geometry, 2. Print. With Corr edn. The MIT Press, Cambridge (1972)
30. Mori, M., MacDorman, K., Kageki, N.: The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* **19**(2), 98–100 (1970)
31. Neth, D., Martinez, A.M.: Emotion perception in emotionless face images suggests a norm-based representation. *J. Vis.* **9**(1), 5–5 (2009)
32. Pazda, A.D., Thorstenson, C.A., Elliot, A.J., Perrett, D.I.: Women's facial redness increases their perceived attractiveness: mediation through perceived healthiness. *Perception* **45**(7), 739–754 (2016)
33. Plutchik, R.: A General Psychoevolutionary Theory of Emotion. In: *Theories of Emotion*, pp. 3–33. Elsevier, Amsterdam (1980)
34. Rule, N.O.: Perceptions of sexual orientation from minimal cues. *Arch. Sex. Behav.* **46**(1), 129–139 (2017)
35. Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161–1178 (1980)

36. Russell, J.A., Fehr, B.: Relativity in the perception of emotion in facial expressions. *J. Exp. Psychol.: Gen.* **116**(3), 223–237 (1987)
37. Said, C.P., Sebe, N., Todorov, A.: Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion* **9**(2), 260–264 (2009)
38. Saldarini, K.: Postal service tests handwriting recognition system. *Government Executive* (1999)
39. Shah, R., Lewis, M.: Locating the neutral expression in the facial-emotion space. *Vis. Cognit.* **10**(5), 549–566 (2003)
40. Sinha, P., Balas, B., Ostrovsky, Y., Russell, R.: Face recognition by humans: nineteen results all computer vision researchers should know about. *Proc. IEEE* **94**(11), 1948–1962 (2006)
41. Snow, J.: <https://www.aclunc.org/blog/>
42. Tae, J., Nam, Y.e., Lee, Y., Weldon, R.B., Sohn, M.H.: Neutral but not in the middle: cross-cultural comparisons of negative bias of “neutral” emotional stimuli. *Cognit. Emot.* **34**(6), 1171–1182 (2020)
43. Thorstenson, C.A., Pazda, A.D., Elliot, A.J., Perrett, D.I.: Facial redness increases men’s perceived healthiness and attractiveness. *Perception* **46**(6), 650–664 (2017)
44. Tronick, E., Als, H., Adamson, L., Wise, S., Brazelton, T.B.: The infant’s response to entrapment between contradictory messages in face-to-face interaction. *J. Am. Acad. Child Psychiatry* **17**(1), 1–13 (1978)
45. Tsankova, E., Kappas, A.: Facial skin smoothness as an indicator of perceived trustworthiness and related traits. *Perception* **45**(4), 400–408 (2016)
46. Tskhay, K.O., Rule, N.O.: Emotions facilitate the communication of ambiguous group memberships. *Emotion* **15**(6), 812–826 (2015)
47. Yip, A., Sinha, P.: Role of color in face recognition. *J. Vis.* **2**(7), 596–596 (2010)
48. Zebrowitz, L.A., Kikuchi, M., Fellous, J.M.: Facial resemblance to emotions: group differences, impression effects, and race stereotypes. *J. Pers. Soc. Psychol.* **98**(2), 175–189 (2010)

Chapter 10

Multi-Stream Temporal Networks for Emotion Recognition in Children and in the Wild



Panagiotis P. Filntisis, Niki Efthymiou, Gerasimos Potamianos, and Petros Maragos

Abstract In this chapter, we extend and leverage the temporal segment networks framework for emotion recognition in children and in the wild. To that end, we explore the effect of different information streams (Body, Face, Context, Audio, Word Embeddings) and representations (RGB, Flow). We perform an extensive ablation analysis, including the effect of each representation and modality on different emotions, and verify the performance of the proposed systems against the previous SoTA methods in the EmoReact and the BoLD datasets.

10.1 Introduction

Automatic human affect recognition from visual cues is an important area of computer vision that has attracted increased interest over the last two decades, due to its many applications. Indeed, social robotics [8], psychiatric care [21], and edutainment [18] are all areas that can benefit from automatic recognition of emotion.

Most past approaches to the problem have focused on facial expressions in order to determine the emotional state of the person of interest [14, 29, 33]. This is reasonable due to the fact that facial expressions have been studied extensively in the psychology and emotion literature [15]. For example, the Facial Action Coding System (FACS) [16] identifies the units of facial movements, based on facial muscle groups. Combinations of the so-called action units (AUs) have also been linked with emotional states with extensions of the basic FACS such as EMFACS (Emotion

P. P. Filntisis (✉) · N. Efthymiou · P. Maragos

School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece

e-mail: filby@central.ntua.gr; [neftymiou@central.ntua.gr](mailto:nefthymiou@central.ntua.gr); maragos@cs.ntua.gr

G. Potamianos

Department of Electrical and Computer Engineering, University of Thessaly, Volos, Greece

e-mail: gpotam@ieee.org

FACS) [19]. On the other hand, there is no similar established coding system for body expressions, although some have been proposed [9].

Compared to facial expression based approaches, recent works have sought alternative modalities and streams of information to detect emotion; one is bodily expressions since many have highlighted the fact that the emotional state is conveyed through bodily expressions as well, and in certain emotions it is the main modality [11, 25, 40], or can be used to correctly disambiguate the corresponding facial expression [2]. Simultaneously, it is important to note that in cases and applications where the emotion needs to be identified, the human body is more frequently available than the face since the face can be occluded, hidden, or far in the distance. Another auxiliary stream of information besides the face and the body that can help in identifying emotions is the context and the surrounding environment of the person [26, 32]. It is apparent that both the place, as well as objects and other humans can influence a person’s emotions. Finally, a different stream of information that can be used on parallel with the visual one is also human speech [17]. It has been shown in many studies that audiovisual fusion can boost emotion recognition performance compared to simply relying on the visual stream [1, 3]. We should also note that inherently emotion recognition is a multi-label problem—the subject might be feeling two or more emotions. This is true, especially when considering an extended set of emotions, as in [30]. The emotions in extended sets do not have the same “semantic” distance between them. For example, anger is more close to annoyance than to happiness. Considering that previous works have showed the superiority of methods that attempt to learn a joint embedding space that contains both word embeddings and visual representations [13, 20, 38, 43], we believe that trying to attach a semantic meaning to the extracted visual feature is a natural way forward.

Based on the aforementioned, in this chapter, we explore the effect of different information streams (Body, Face, Context, Audio, Word Embeddings) and representations (RGB, Flow) in two different emotion recognition applications: the first one is emotion recognition in the wild, and the second one emotion recognition in children. We do this by building two different multi-stream emotion recognition architectures for the two applications, each one leveraging a different set of information streams. The first multi-stream architecture for emotion recognition in the wild focuses on body language, contextual information and semantic embeddings of labels, while the second multi-stream architecture for emotion recognition in children leverages the audio modality, alongside facial expressions. Both architectures also take advantage of both RGB and Flow representations in order to boost performance.

The rest of the chapter is organized as follows: Sect. 10.2 discusses related work on emotion recognition. Section 10.3 describes in detail our two proposed multi-stream emotion recognition architectures, and Sect. 10.4 presents our thorough experimental results on the BoLD and EmoReact datasets. Finally, Sect. 10.5 provides our conclusions and directions for future work.

10.2 Related Work

While most past approaches in visual detection of affect have been focused on facial expressions [11], recent approaches have started taking into account the body language [25] of the person in question, as well as its surrounding context/environment.

In [23], Gunes and Piccardi introduced a bimodal architecture that takes into account both upper body and facial expressions, in order to detect affect in videos. In [10], Dael et al. analyzed and classified body emotional expressions using a body action and posture coding system which was proposed in [9]. The 3D pose of children was also utilized in [31] by Marinou et al. to detect emotions in continuous dimensions, while in [18], 2D pose was used and fused with facial expressions for child emotion recognition. Luo et al. [30] introduced a large scale video dataset (BoLD) annotated with categorical and continuous emotions, which is the one used in the Bodily Expressed Emotion Understanding (BEEU) challenge.

Regarding the context modality, Kosti et al. [26] introduced a large scale dataset for emotion recognition (EMOTIC) in different contexts (e.g., other people, places, or objects) and a convolutional neural network (CNN) based two-stream architecture that focused on the body and context of the subjects. The CAER video dataset for context-based emotion recognition was presented in [27], along with a two-stream architecture which employed adaptive-fusion to merge the two streams. In [32], Mittal et al. designed a deep architecture with several branches, focusing on different interpretations of the surrounding context (e.g., environment and interaction context) to significantly increase resulting predictions in the EMOTIC dataset.

However, although a number of studies have indeed emphasized the importance of leveraging multiple modalities for emotion recognition in adults [5, 12, 36], there is a lack of works studying multiple modalities for emotion recognition in children.

Regarding facial emotion recognition in children, Goulart et al. proposed in [22] a computational system for estimating children emotion during Child-Robot Interaction (CRI), deploying visual information from both RGB and infrared thermal cameras. The proposed system detects the facial regions of interest that are relevant to five basic emotions. Lopez-Rincon in [28] proposed a Convolutional Neural Network (CNN) combined with a Viola-Jones face detector, trained using the AffectNet database [33], and tuned it with children data in order to recognize children facial emotional expressions. Marinou et al. [31] proposed an automated approach using 3d skeleton data and a CNN architecture for action and continuous emotion recognition during robot-assisted therapy sessions of children with Autism Spectrum Disorders (ASD). In [7], a system perceived children affective expressions while playing chess with an iCat robot and modified the behavior of the robot to be more friendly and increase children engagement.

Apart from the face, which is the most commonly used channel for identifying emotion [11], there are other modalities equally powerful to reveal children affect such as speech and body movements. In [34], an ensemble of AlexNet networks was applied on multiple spectrograms in order to extract deep features, which were then

used by an SVM to identify emotions in the EmoReact dataset. For the same dataset, [35] combined traditional audio features and features extracted from the OpenFace framework [4] (action units, shape parameters and head orientation) with an SVM for audiovisual emotion recognition. In [18], we proposed a two-branch architecture modeling body movements along with the facial expressions to identify emotions in children during CRI scenarios.

Finally, some recent works have also focused on extracting visual representations from images that present the semantic relations found in embeddings built from words. The DeViSE embedding model [20] extracted semantically-meaningful visual representations by introducing a similarity loss between the feature vector extracted from a CNN and the word embedding from a skip-gram text model. Using a similar method, Wei et al. [42] built joint text and visual embeddings as emotion representation from web images, and in [44], Ye and Li built semantic embeddings for a multi-label classification problem.

10.3 Multi-Stream Architectures

In this section we present our two multi-stream architectures for emotion recognition from multiple information streams. Before delving into the specific details of each architecture, we briefly discuss the temporal segment networks framework [41], which constitutes a building block of both our architectures.

10.3.1 Temporal Segment Networks

In the temporal segment networks framework [41], an input video is split into K different segments of equal duration M , and in the next step, a snippet of length $N < M$ consecutive frames is randomly sampled from each segment, resulting in K snippets T_k . Subsequently, each snippet is fed to a CNN, yielding class scores S_k (in our case emotion scores) for each snippet. In the last step, the scores of the different snippets are fused using the segmental consensus function H that is applied on the representations of all different snippets to obtain the final scores:

$$S = H(S_k) = H(F_v(T_k; \mathbf{W}_v) |_{k \in K}) \quad (10.1)$$

where $F_v(T_k; \mathbf{W}_v)$ denotes the application of a CNN with parameters \mathbf{W}_v on the snippet T_k . The most common consensus function that can be used is averaging, while others include maximum or weighted averaging (we use simple averaging). The CNN is then trained using standard cross-entropy loss in the case of multiclass classification, or binary cross-entropy in the case of multilabel classification (which is the case of both emotion recognition use-cases we consider).

Traditionally, TSNs take input from both the RGB of the input video, as well as the optical flow, with each one trained separately and then fused using average or weighted average fusion. As with TSNs for action recognition, we also use both modalities, since the optical flow can be used to model the dynamics that arise during expressions of emotion, while the RGB modality can best identify static expressions such as smiles.

The paradigm of TSNs offers several benefits to emotion recognition. Considering an input video with a person expressing emotion, the archetype facial expressions and action units that correspond to each emotion are not present throughout the video, but usually only during a short period of it. As a result, temporal sampling allows the network to access several parts of the video and model its long-range temporal structure, thus being more likely to observe the corresponding facial expression. In addition, compared to processing the entire video, the sampling process ignores redundant information in consecutive video frames, helping avoid overfitting and offering a type of data augmentation, valuable for databases of small size.

10.3.2 First Multi-Stream Architecture—Emotion Recognition in the Wild

The first architecture is shown in Fig. 10.1. Here, we employ three different streams of information: Body, Context (i.e., scene) and word embedding information from emotion labels.

10.3.2.1 Body and Context

We crop the input video around the body of the person of interest using Open-Pose [6] in both RGB and Flow modalities. We denote these streams as RGB-b and Flow-b. Then, we also introduce one additional stream based on the context-environment surrounding the annotated human. For the RGB modality, we input the context in the network in the same way as in [32], by masking out the instance body (we set all pixels to 0). We call this stream RGB-c. During training, the RGB-b and RGB-c streams are combined at the feature level (RGB-bc) and are trained jointly while the Flow-b TSN is trained independently.

10.3.2.2 Embedding Loss

The second stream of information is input in the network through the inserting of an embedding loss on the feature vector extracted by the Convolutional Neural Network (ConvNet). This is done to exploit the fact that some emotions are closer semantically to others. This is also revealed by examining the correlation matrix of

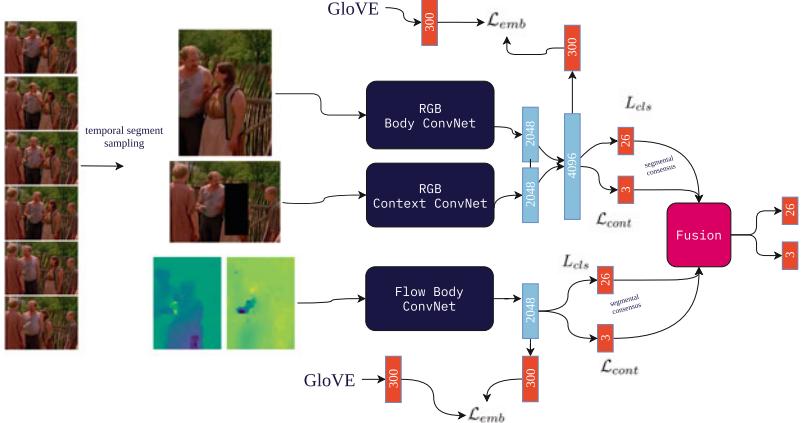


Fig. 10.1 First multi-stream architecture employing TSNs with two RGB spatial streams (body and context) and one optical flow stream. The final results are obtained using average score fusion

the dataset labels in [30], where some labels occur more frequently in combination with others (e.g. Happiness and Pleasure, Annoyance and Anger, etc.). Due to this result, we try to attach a semantic meaning to the feature vector extracted by the backbone image network.

To implement this, we first obtain for each one of the 26 categorical labels of BoLD [30] their 300-dimensional GloVE word embedding [37]. A PCA-projection of the 26 embeddings is shown in Fig. 10.2, where it is apparent that the distances between embeddings are indicative of their “semantic” distance. We then use a fully connected layer to map the feature extracted from the image to a 300-dimensional space and introduce the following mean-squared based loss:

$$\mathcal{L}_{emb} = \|\mathbf{W} f_v(\mathbf{x}) - \frac{1}{|K|} \sum_{y \in K} f_w(y)\|_2 \quad (10.2)$$

where $f_v(\mathbf{x})$ is the feature vector extracted by applying the convNet on the image \mathbf{x} , \mathbf{W} is a linear transformation from the space of the feature vector to the word embedding space, $f_w(y)$ is the word embedding of the label y , and K is the set of all positive labels for the image \mathbf{x} . That is, we try to reduce the Euclidean distance between the projected image feature and the arithmetic mean of the GloVE embeddings of the positive labels for image/video.

10.3.2.3 Predictions

Finally, after extracting for each sampled image its feature vector, we use two fully connected layers, one to classify to the 26 different categorical labels, and one to

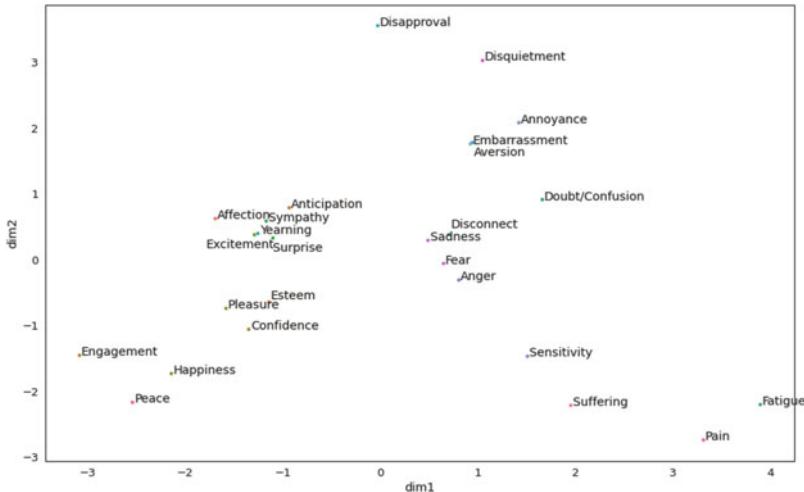


Fig. 10.2 PCA projection of the categorical emotions GloVe word embeddings

regress over the 3 different categorical emotions. The two TSNs are trained using the following loss:

$$\mathcal{L} = \mathcal{L}_{cls_1} + \mathcal{L}_{cls_2} + \mathcal{L}_{cont} + \mathcal{L}_{emb} \quad (10.3)$$

Specifically, since the dataset does not provide explicitly the multilabel targets, but the crowdsourced scores between 0 and 1, we include two different losses for the classification part: \mathcal{L}_{cls_1} that is the binary cross-entropy between the predicted scores and the multilabel target (obtained after thresholding the multilabel scores at 0.5) and \mathcal{L}_{cls_2} that is the mean squared error between the predicted scores and the multilabel scores. We empirically found that the inclusion of \mathcal{L}_{cls_2} slightly boosted performance. For the regression part, \mathcal{L}_{cont} is the mean-squared error between the regressed values and the continuous emotions. Finally \mathcal{L}_{emb} is as in (10.2).

10.3.3 Second Multi Stream Architecture—Child Emotion Recognition

The second multi-stream architecture we introduce leverages two different information streams (Fig. 10.3) and is built for child emotion recognition.

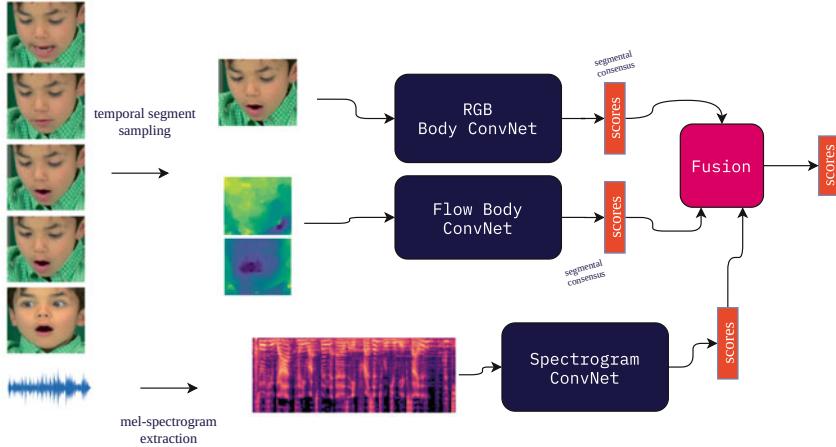


Fig. 10.3 The proposed multimodal emotion recognition architecture for child-robot interaction

10.3.3.1 Face

For the face stream, we crop the input video (both RGB and Flow) around the face of the person, by using the facial landmarks obtained by OpenFace [4]. The rest of the training procedure is identical with the first architecture.

10.3.3.2 Audio Branch

In the audio branch, considering the input waveform of the video, we first extract its mel-spectrogram representation and then apply a CNN $F_a(W_a)$ on it in order to extract the audio representation. Here, we bypass the cumbersome feature extraction methods by considering the mel-spectrogram of the waveform as an image, and applying standard computer vision techniques. Next, as with the visual modality, a fully connected layer is used in order to obtain the final emotion scores. The audio branch is susceptible to overfitting because the full spectrogram is fed to the network, contrary to the visual branch where temporal sampling is used. To counter this, we apply a more aggressive regularization scheme with high penalty for L2 regularization during training.

10.3.3.3 Training and Audiovisual Fusion

In order to fuse information from the visual and audio modalities, we consider two different types of fusion between both RGB-audio, as well as Flow-audio modalities: feature fusion and score fusion, and two training schemes: independent training and joint training.

During joint training, the RGB (or Flow) and audio CNN are trained concurrently, and depending of the fusion scheme, we either concatenate their feature vectors (feature fusion) before the last fully connected layer, or average the scores (score fusion) obtained after the last fully connected layer. In order to achieve feature fusion under joint training, we repeat the audio feature vector K times (where K is the number of segments/snippets), and associate each visual snippet with the audio feature vector for the whole video, through concatenation of the feature vectors. In contrast, in independent training the RGB (or Flow) and audio networks are trained separately, and we then average their emotion scores.

10.4 Experimental Results

10.4.1 First Use-case: Multi-Stream Emotion Recognition In-the-wild

We evaluate the first multi-stream architecture on the BoLD (Body Language Dataset) corpus [30] consisting of 9876 video clips of humans expressing emotion, primarily through body movements. Each clip can contain multiple characters, yielding a total of 13,239 annotations, split into a training, validation, and test set. The dataset has been annotated by crowdsourcing employing two widely accepted categorizations of emotion. The first one is the categorical annotation with a total of 26 labels first used in [26], by collecting and processing an extensive affective vocabulary. The second annotation regards the continuous emotional dimensions of the VAD (Valence–Arousal–Dominance) Emotional State Model [39]. The metric we use for evaluation is the following Emotion Recognition Score (ERS) [30]:

$$\text{ERS} = \frac{1}{2} \left(mR^2 + \frac{1}{2}(mAP + mRA) \right) \quad (10.4)$$

where mR^2 is the mean coefficient of determination (R^2) score for the three dimensional emotions (VAD), and mAP and mRA is the mean Average Precision and the mean area under receiver operating characteristic curve (ROC AUC) of the multilabel categorical predictions.

We train each TSN for 50 epochs using Stochastic Gradient Descent (SGD), with initial learning rate 10^{-3} which drops by a factor of 10 at 20 epochs.¹ The backbone networks used is a residual network (ResNet) with 101 layers for the body convNets and a ResNet with 50 layers for the context convNet. We use the default hyperparameters of TSNs: 3 segments, 1 frame from each segment for the RGB streams, and 5 frames from each segment for the optical flow stream. The

¹ PyTorch code available at <https://github.com/filby89/NTUA-BEEU-eccv2020>.

Table 10.1 Ablation experiment by training with and without \mathcal{L}_{emb}

	Model	<i>mAP</i>	<i>mRA</i>	<i>mR²</i>	<i>ERS</i>
Without \mathcal{L}_{emb}	RGB-b	0.1567	0.6140	0.0538	0.21955
	Flow-b	0.1444	0.5914	0.0507	0.2093
	RGB-b + Flow-b	0.1623	0.6307	0.078	0.2375
With \mathcal{L}_{emb}	RGB-b	0.1564	0.6143	0.0546	0.21997
	Flow-b	0.1465	0.5947	0.0579	0.2142
	RGB-b + Flow-b	0.1637	0.6327	0.0874	0.2428

Bold values denote the best result

Table 10.2 Results on the validation and test set of BoLD including the RGB context stream and \mathcal{L}_{emb}

Set	Model	<i>mAP</i>	<i>mRA</i>	<i>mR²</i>	<i>ERS</i>
Valid	RGB-c	0.1395	0.5760	0.0365	0.1971
	RGB-bc	0.1566	0.6055	0.0675	0.2243
	RGB-bc + Flow-b	0.1656	0.6266	0.0917	0.2439
Test	RGB-bc + Flow-b	0.1796	0.6416	0.1141	0.26235

Bold values denote the best result

consensus used for segment fusion is averaging. For each network, we select the epoch with the best validation ERS. We have also found experimentally that the partialBN (Batch Normalization) technique used in [41] gives a nontrivial boost to the performance of the network.

First, in Table 10.1 we present two ablation experiments regarding the addition of \mathcal{L}_{emb} . We can see that adding the embedding loss increases slightly the performance in the RGB-b stream, and gives a boost to the performance of the Flow-b stream.

Then, in Table 10.2 we present our experimental results on the validation set of BoLD including the RGB context stream. From the results we can see that including the context along with the body in the RGB modality boosts the validation ERS of the architecture. We also experimented with including the context in the Flow network, but this resulted in worse performance. Our final result for the test set was the model with the best validation score (0.2439 employing RGB-bc + Flow-b), using 25 segments instead of 3. The results of the different metrics on the test set can also be seen in Table 10.2, while the final ERS is 0.26235, improving upon the previous best result of 0.2530 [30].

10.4.2 Second Use-Case: Child Emotion Recognition

We evaluate the second multi-stream architecture for child emotion recognition on the EmoReact dataset. The EmoReact dataset [35] contains videos of 63 children (32F, 31M, aged 4 to 14) reactions to different topics, and has been collected from the YouTube channel React. The number of all videos across the training (432

videos), validation (303 videos), and test set (367) is 1102. Each video is annotated with one or more emotions, from a total of 8 emotion labels: Curiosity, Uncertainty, Excitement, Happiness, Surprise, Disgust, Fear, and Frustration. To the best of our knowledge, the EmoReact dataset is the only dataset of children expressing emotion, both verbally and visually.

Like in the first architecture, the CNN backbone of the visual and audio branches is a residual CNN with 50 layers (ResNet50) [24]. Specifically for the CNN of the visual RGB branch, we have pretrained it on the largest facial expression dataset, AffectNet [33], achieving 59.47% accuracy on the validation set (test set is not available). Because the label distribution of AffectNet is highly skewed, we employ balanced sampling so that the network sees the underrepresented classes more often. The residual networks of the audio branch and Flow modality are pretrained on ImageNet (we obtain the weights of the network as provided by the PyTorch framework).

We train all networks and modalities with stochastic gradient descent for 60 epochs, starting with a learning rate of $1e-2$, momentum 0.9, and regularization with weight decay (L2 regularization) $5e-4$. The learning rate is reduced by a factor of 10 at 20 and 40 epoch milestones.² Training is done using binary cross-entropy loss. For evaluation, we select the epoch with the best validation area under receiver operating characteristic (ROC AUC), and apply the corresponding network on the test set, reporting class-balanced and unbalanced ROC AUC. Especially in the case of audio, we found out that a more aggressive regularization scheme is needed to avoid overfitting, and thus we increased the weight decay ten-fold to $5e-3$.

10.4.2.1 Number of Segments

As a first ablation study, we consider the number of segments (and as a consequence the number of snippets), which are used during training of the visual branch with the RGB and Flow modalities. We consider 4 different values: 1, 3, 5, and 10, and report in Table 10.3 the results on the ROC AUC (balanced per class and unbalanced), as well as average time taken per epoch for training and inference, on a computer with an RTX 2080 GPU.

We can see that in the case of RGB, increasing the number of segments above 3 does not result in significant performance difference, showing that even a small number of segments can achieve satisfactory performance. However, increasing the number of segments increases significantly both the training and inference times. For the Flow modality, we see that selecting 5 as a number of segments results in a balanced trade-off between performance and execution time, since the performance increase using 10 segments is minuscule. For the following experiments, we use 3 segments for RGB and 5 segments for the Flow modality.

² We have made the code for the experiments publicly available at <https://github.com/filby89/multimodal-emotion-recognition>.

Table 10.3 ROC AUC and average time elapsed per epoch with varying number of sampled snippets

Segments	ROC AUC		Sec/train epoch	Sec/val epoch
	Balanced	Unbalanced		
RGB				
1	0.685	0.773	11	7
3	0.713	0.786	27	20
5	0.709	0.787	40	26
10	0.715	0.788	73	51
Flow				
1	0.585	0.741	37	23
3	0.596	0.744	101	70
5	0.623	0.757	166	115
10	0.627	0.759	294	210

Table 10.4 Results on the EmoReact dataset for different fusion and training schemes between the RGB-audio and Flow-audio modalities

Fusion	Training	ROC AUC	
		Balanced	Unbalanced
Single Modality	Audio	0.715	0.750
	Visual (RGB)	0.713	0.786
	Visual (Flow)	0.623	0.757
Score Fusion RGB-audio	Joint Training	0.720	0.756
	Independent Training	0.747	0.799
Score Fusion Flow-audio	Joint Training	0.719	0.746
	Independent Training	0.725	0.787
Feature Fusion RGB-audio	Joint Training	0.719	0.769
Feature Fusion Flow-audio	Joint Training	0.707	0.744

Bold values denote the best result

10.4.2.2 Audiovisual Fusion and Training Schemes

Next, we experiment with the different kinds of fusion schemes that can be used to merge the RGB and audio, as well as the Flow and audio modalities: feature vs. score fusion, as well as the pretraining scheme: joint training of both networks vs. independent training. The results of this study are shown in Table 10.4. Training the networks independently and then averaging their scores achieves the best result in both cases of audiovisual fusion (RGB-audio and Flow-audio), when compared to both the single modalities, as well as their fusion using joint training. This could be attributed to the fact that while the TSN framework inherently avoids overfitting using the temporal sampling, in the case of audio this is not the case, since the full spectrogram is used, and more elaborate schemes of regularization are needed.

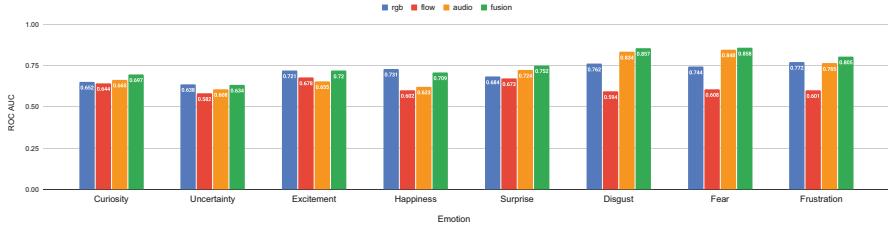


Fig. 10.4 ROC AUC per emotion, for each different modality and their average score fusion

Table 10.5 Final ROC AUC results on the EmoReact dataset

	ROC AUC	
	Balanced	Unbalanced
Audio		
Audio features + SVM [35]	0.610	–
dnn ensemble + SVM [34]	0.718	–
Ours (End-to-End)	0.715	0.750
Visual		
Openface + SVM [35]	0.620	–
Ours (Flow)	0.623	0.757
Ours (RGB)	0.713	0.786
AudioVisual		
[35]	0.640	–
Ours (RGB+Audio+Flow)	0.754	0.809

Bold values denote the best result

10.4.2.3 Emotion by Modality

Next, we explore the strengths and weaknesses of each different modality, by showing the different ROC AUC scores for each emotion, in Fig. 10.4. We observe that especially for Happiness, RGB is the most appropriate modality, while Fear and Disgust, are best identified through the children’s speech. Flow, in almost all cases underperforms when compared to the other modalities, however in the case of Excitement and Surprise it achieves a high score, which can be explained by the more intense movements a person does when expressing these emotions. The figure also shows the result of average score fusion using independent training for all three modalities, RGB, Flow, and audio. We can see that overall, fusion increases the total balanced and unbalanced scores, however in the case of Uncertainty, Excitement, and Happiness, it results in slightly lower score when compared to RGB only.

10.4.2.4 Final Results

We present the final results of the emotion recognition system on EmoReact in Table 10.5, where we have also added the result of average score fusion between

the three different modalities (using independent training), as well as the previous reported best results in the literature. For the audio modality, our architecture achieves significantly better ROC AUC than [35], which used a carefully selected speech features set with an SVM, as well as similar results with Nagarajan et al. [34]. However, our approach is end-to-end and simple to implement, while Nagarajan et al. employed an elaborate scheme involving multiple AlexNet architectures for feature extraction and an SVM on top of them to achieve the final result.

In the visual modality, our RGB TSN architecture improves significantly upon the best previous published result, which used features extracted from the OpenFace framework with an SVM [35].

Finally, our audiovisual fusion scheme using all three modalities with independent training further increases the ROC-AUC up to 0.754, resulting in significant score improvement upon all previous studies.

10.5 Conclusions and Future Work

In this chapter, we have explored multi-stream architectures for two well-known emotion recognition applications: emotion recognition in the wild, and emotion recognition in children. To that end, we have designed two different multi-stream TSN-based architecture which leverage various information sources: face, body, context, word embeddings, and audio, as well as visual representations: RGB and Flow. Our architectures have achieved state-of-the-art performance on the BoLD and EmoReact dataset verifying the fact that moving forward, emotion recognition systems should not only focus on facial expressions, but also take into account other information streams. In the future we aim to build a unifying multi-stream architecture capable of processing all different information streams and deploy it for various emotion recognition applications.

References

1. Antoniadis, P., Pikoulis, I., Filntsis, P.P., Maragos, P.: An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 3645–3651 (2021)
2. Aviezer, H., Trope, Y., Todorov, A.: Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* **338**(6111), 1225–1229 (2012)
3. Avots, E., Sapiński, T., Bachmann, M., Kamińska, D.: Audiovisual emotion recognition in wild. *Mach. Vis. Appl.* **30**(5), 975–985 (2019)
4. Baltrušaitis, T., Zadeh, A., Lim, Y.C., Morency, L.: Openface 2.0: facial behavior analysis toolkit. In: Proc. FG, pp. 59–66 (2018). <https://doi.org/10.1109/FG.2018.00019>
5. Bänziger, T., Pirker, H., Scherer, K.: GEMEP-Geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions. In: Proc. LREC, vol. 6, pp. 15–19 (2006)
6. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 172–186 (2019)

7. Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., Mcowan, P.W.: Multimodal affect modeling and recognition for empathic robot companions. *Int. J. Humanoid Rob.* **10**, 1350010 (2013)
8. Cavallo, F., Semeraro, F., Fiorini, L., Magyar, G., Sinčák, P., Dario, P.: Emotion modelling for social robotics applications: a review. *J. Bionic Eng.* **15**(2), 185–203 (2018)
9. Dael, N., Mortillaro, M., Scherer, K.R.: The body action and posture coding system (BAP): development and reliability. *J. Nonverbal Behav.* **36**(2), 97–121 (2012)
10. Dael, N., Mortillaro, M., Scherer, K.R.: Emotion expression in body action and posture. *Emotion* **12**(5), 1085 (2012)
11. De Gelder, B.: Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* **364**(1535), 3475–3484 (2009)
12. De Silva, L.C.: Audiovisual emotion recognition. In: Proc. Int. Conf. on Systems, Man and Cybernetics (2004)
13. Dong, J., Li, X., Snoek, C.G.: Word2visualvec: image and video to sentence matching by visual feature prediction. arXiv preprint arXiv:1604.06838 (2016)
14. Du, S., Tao, Y., Martinez, A.M.: Compound facial expressions of emotion. *Proc. Natl. Acad. Sci.* **111**(15), E1454–E1462 (2014)
15. Ekman, P., Keltnner, D.: Universal facial expressions of emotion. In: Segerstrale, U., Molnar, P. (eds.) *Nonverbal Communication: Where Nature Meets Culture*, pp. 27–46. Routledge, Milton Park (1997)
16. Ekman, R.: *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, Oxford (1997)
17. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognit.* **44**(3), 572–587 (2011)
18. Filntisis, P.P., Efthymiou, N., Koutras, P., Potamianos, G., Maragos, P.: Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction. *IEEE Rob. Autom. Lett.* **4**(4), 4011–4018 (2019)
19. Friesen, W.V., Ekman, P., et al.: Emfacst-7: emotional facial action coding system. Unpublished manuscript, University of California at San Francisco **2**(36), 1 (1983)
20. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: a deep visual-semantic embedding model. In: *Advances in Neural Information Processing Systems*, pp. 2121–2129 (2013)
21. Gaudelus, B., Virgile, J., Geliot, S., Franck, N., Dupuis, M., Hochard, C., Josserand, A., Koubichkine, A., Lambert, T., Perez, M., et al.: Improving facial emotion recognition in schizophrenia: a controlled study comparing specific and attentional focused cognitive remediation. *Front. Psychiatry* **7**, 105 (2016)
22. Goulart, C., Valadão, C., Delisle-Rodriguez, D., Funayama, D., Favarato, A., Baldo, G., Binotte, V., Caldeira, E., Bastos-Filho, T.: Visual and thermal image processing for facial specific landmark detection to infer emotions in a child–robot interaction. *Sensors* **19**, 2844 (2019)
23. Gunes, H., Piccardi, M.: A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: Proc. ICPR, vol. 1, pp. 1148–1153 (2006)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
25. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: a survey. *IEEE Trans. Affect. Comput.* **4**(1), 15–33 (2013)
26. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotion recognition in context. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1960–1968 (2017)
27. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. In: Proc. IEEE International Conference on Computer Vision, pp. 10143–10152 (2019)
28. Lopez-Rincon, A.: Emotion recognition using facial expressions in children using the NAO robot. In: Proc. CONIELECOMP, pp. 146–153 (2019)

29. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 94–101 (2010)
30. Luo, Y., Ye, J., Adams Jr., R.B., Li, J., Newman, M.G., Wang, J.Z.: ARBEE: towards automated recognition of bodily expression of emotion in the wild. *Int. J. Comput. Vis.* **128**(1), 1–25 (2020)
31. Marinou, E., Zanfir, M., Olaru, V., Sminchisescu, C.: 3D human sensing, action and emotion recognition in robot assisted therapy of children with autism. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2158–2167 (2018)
32. Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emoticon: context-aware multimodal emotion recognition using Frege’s principle. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14234–14243 (2020)
33. Mollahosseini, A., Hasani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**(1), 18–31 (2017)
34. Nagarajan, B., Oruganti, V.R.: Cross-domain transfer learning for complex emotion recognition. In: Proc. TENSYMP (2019)
35. Nojavanaghari, B., Baltrušaitis, T., Hughes, C.E., Morency, L.P.: EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In: Proc. ICMI (2016)
36. Pantic, M., Sebe, N., Cohn, J.F., Huang, T.: Affective multimodal human-computer interaction. In: Proc. Int. Conf. on Multimedia (2005)
37. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
38. Ren, Z., Jin, H., Lin, Z., Fang, C., Yuille, A.L.: Multiple instance visual-semantic embedding. In: Proc. BMVC (2017)
39. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. *J. Res. Pers.* **11**(3), 273–294 (1977)
40. Tracy, J.L., Robins, R.W.: Show your pride: evidence for a discrete emotion expression. *Psychol. Sci.* **15**(3), 194–197 (2004)
41. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: towards good practices for deep action recognition. In: European Conference on Computer Vision, pp. 20–36. Springer, Berlin (2016)
42. Wei, Z., Zhang, J., Lin, Z., Lee, J.Y., Balasubramanian, N., Hoai, M., Samaras, D.: Learning visual emotion representations from web data. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13106–13115 (2020)
43. Wortman, B., Wang, J.Z.: Hicem: a high-coverage emotion model for artificial emotional intelligence. arXiv preprint arXiv:2206.07593 (2022)
44. Yeh, M.C., Li, Y.N.: Multilabel deep visual-semantic embedding. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(6), 1530–1536 (2020)

Part IV

Photography, Arts

This part explores ways that art and technology intersect and interact. Each chapter investigates a unique perspective of this relationship.

The first chapter, “The Formal Language of Photography: A Primer,” unravels the perplexing subject of the language of photography. It deconstructs the visual medium into its essential elements—tones, colors, lines, shapes, and textures—and examines how they elicit emotional responses.

The next chapter, “Breathing With Robots: Notating Performer Strategy, Alongside Choreographer Intent and Audience Observation, In Breath-driven Robotic Dance Performance,” the interaction between a breath-activated robot and a dancer is explored, dissecting the intertwined elements of visual aesthetic, style, and emotion.

The last chapter, “Humanist-in-the-Loop: Machine Learning and the Analysis of Style in the Visual Arts,” probes into the challenges of using computational models for artistic style analysis, advocating for a contextual approach that blends computer science and art history.

Together, these chapters represent an engaging and thought-provoking journey through the entwined union of art and technology, pushing the boundaries of how we perceive, create, and analyze art.

Chapter 11

The Formal Language of Photography: A Primer



QT Luong

Abstract Photography works as a language on several levels. At an abstract level, the image is viewed as a set of formal elements rather than signifiers representing objects from the real world. This chapter, from the point of view of a professional photographer, breaks down the underlying structure of photographs into simple elements and enumerates the emotional responses attached to them, generally along the fundamental polarities of harmony and tension. Those elements are primarily tones and colors, from which are derived the secondary and more abstract elements of line, shape, and texture. As a result of how cameras transform the world, besides representing three-dimensional space, all photographs share common qualities specific to the medium: vantage point, frame, time, and focus. If all those elements are the words of the visual language of photography, then composition principles such as contrast, balance, unity, and emphasis are its syntactic rules. A formal analysis of a photograph depends on minimum domain knowledge, however, a complete account of how humans read photographs should also involve the world being photographed, the photographer, and society.

11.1 Introduction

Our modern culture has seen an accelerated move from word-dominated to visually-dominated communications, and the catalyst of this change has been photography. As a form of communication, photography has its own language. There are four functions in critical analysis: describing, interpreting, evaluating, and theorizing [1]. In examining the visual language of photography, this chapter is concerned mostly with the first function, a prerequisite for the others. It is hoped that learning how a practitioner of photography reads photographs can help provide a roadmap for those trying to build automated systems to analyze them.

QT Luong (✉)
Terra Galleria Photography, San Jose, CA, USA
e-mail: qtl@terragalleria.com

Each visual message exists at three overlapping levels [4]. The representational level consists of the objects, life, and facts that we can identify from our experience of the world. The abstract level is driven by our instinctual response to the elemental visual components of the message. The symbolic level involves the decoding of a system to which humans have arbitrary attached meaning. Those levels approximately parallel the semiotic categories of the indexical, the iconic, and the symbolic [8].

It takes great skill to produce life-like paintings. When photography was invented in the nineteenth century, it instantly became the most reliable method for representing visual reality. Representation is natural to the camera. At the first level, which is fundamental [12], photographs are read as a substitute reality, an illusion of a window into the world, so much that laymen frequently conflagrate them with the real thing that they depict, the subject matter, or contents. But sometimes, the purpose of the photograph does not rely so much on its contents as on the way it is presented.

Embedded in the previous level is a second level, abstract in quality. Containing signals to our perceptual system, it consists of the understructure of the photograph: its formal elements. They are an abstraction of the contents in the sense that extraneous details are superseded, emphasizing distinctive features. Abstraction conveys an essential, pre-conceptual, pre-interpretive meaning emotionally, bypassing the conscious to reach directly the unconscious.

Students of photography understand that much of the reality depicted in pictures is filtered and altered through the artistic choices made by the photographer, how the photograph is composed and its elements arranged. The photographer, by selecting a composition from all possible compositions, encodes a scene with his or her interpretation, and the viewer decodes it by reading the photograph. Form is how the subject matter is presented. Even though the objects being photographed are recognized as representing real objects, once they are photographed, their representations acquire a new significance as design elements.

Making sense of the representational and symbolic levels requires a comprehensive knowledge base, as well as an accounting of the photographer and viewer's particular experiences, contexts, and biases. On the other hand, making sense of the abstract level only requires identifying its formal elements. In that reading, they consist of no more than two-dimensional markings, rather than signifiers of things in the world. Photographs are visual communication devices that use a specific language, which like all languages, has its own words and syntactic rules. It is possible to analyze that structure and break it down into elements. An informal, empirical, but rather widespread tradition associates each of those element attributes with a range of emotional responses. Although specific to each element, they often align with the polarities of harmony and tension that are so fundamental to human perception. As the meaning of photographs containing human figures or cultural elements is often directly derived from their subject matter, this chapter uses as examples nature landscape photographs made by the author in America's parks to illustrate the relationship between the photograph's emotional content and its visual language.

11.2 Primary Visual Elements

Photographs are a recording not of objects, but the light reflected off them. Although they are eventually made of discrete elements such as pixels, grain, or pigments, the way they are normally viewed, those elements are not noticed because they are small. When we say that an image is “pixelated” or “grainy”, we may imply a low production quality unless we recognize intentionality. The primary component of photographs is the light recorded at each point of the picture plane. From a perceptual point of view, it is convenient to describe it in terms of luminance, hue, and saturation. Luminance coincides with the photographic quality of tone. In black and white images, tone is the same as shade of greys. The hue and saturation together define what is generally referred to as color.

Even before identifying the elements of a photograph, we can perceive its overall tonality and color palette, even in a low-resolution version. The tones and colors in a photograph are a result of a complex imaging process that involves physical factors—the direction, apparent size, brightness, and color of the light source and the surface of the subject, and technical factors—the exposure and any post-processing. Automatic exposure has been available in cameras for a long time, but the photographer could deliberately make the photograph lighter or darker to convey his vision. In current imaging systems such as camera phones that incorporate computational photography, automation goes a step further, as a massive amount of image processing takes place before the image is displayed to the user. Such processing involves operations that can modify not just exposure, but also all the primary visual element attributes discussed in this section, including local contrast and color balance. While most users are content with the resulting clearer and more vibrant images, the same systems usually provide post-processing tools that let users regain some control, should it be desired.

11.2.1 *Tone and Color*

Variations in tone are what allow humans to visually make sense of their complex environment in a way that is sufficient for survival. They are so dominant in importance for content identification that viewers readily accept a monochromatic representation of reality such as a black and white photograph. Color has less practical value for identification, but carries a strong symbolic and emotional message.

Does the mere choice to photograph in black and white signify something? The use of a monochromatic palette is rare in painting, where it is associated with explorations of fantasies and the psyche, such as in the work of Max Klinger and Odilon Redon. However, so many influential photographs are monochromatic that it has been argued that black and white is the language of photography. Although color photography became practical in the mid-1930s, it was not widely accepted in the



Fig. 11.1 High/low key. *Pacific Rim National Park, British Columbia, 2004. Lake McDonald, Glacier National Park, Montana, 2015.* All photos in this and other figures of this chapter are from the author, QT Luong. More information is at terragalleria.com

art world until the 1970s [14], and some practitioners continue to prefer black and white photography. They see it as a tonal language specific to photography and in particular to the documentary tradition - although this could be a Western preference [11]. Black and white photography emphasizes expression, stylization, distancing, and abstraction, whereas in their view color is distracting and vulgar. On the other hand, color adds a new level of descriptive information and specificity. It makes the experience of viewing the photograph more transparent and immediate.

Color also adds a new level of expression, but only if the photograph is designed around color, not if it is a photograph that happens to be in color [14]. For instance, color can be used as a compositional element, or a color palette can be selected to evoke a mood or communicate an idea.

11.2.2 Tonality

An overall dark image (“**Low key**”) with mostly mid-tones and dark tones evokes a dark, ominous, and mysterious mood, whereas an overall light image (“**High key**”) with mostly mid-tones and light tones evokes a light, optimistic, and pure mood (Fig. 11.1). In Western culture, black connotes death when dull and formality when glossy, whereas white is associated with purity and cleanliness.

The other qualities related to tonality that can be perceived at a glance are how many tones the photograph contains (“Tonal gradation”), and their degree of separation (“Tonal contrast”). They are eminently characteristic qualities of photographs. An image with **high tonal gradation** contains a smooth tonal scale with a full range of tones that are perceived as organic and realistic. An image with **low tonal gradation** has only a few tones, and because of that appears abstracted from reality. High tonal gradation also lends to the perception of a high level of detail and depth (Fig. 11.2).



Fig. 11.2 Low/high tonal gradation. *Willow Flats, Grand Teton National Park, Wyoming, 2011.* *Siesta Lake, Yosemite National Park, California, 2001*

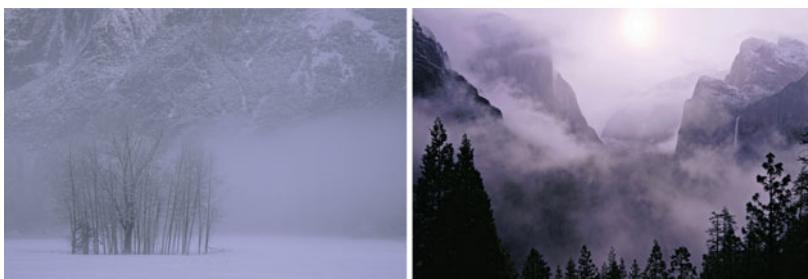


Fig. 11.3 Low/high global tonal contrast. *Cook Meadow, Yosemite National Park, California, 1999.* *Tunnel View, Yosemite National Park, California, 1998*

11.2.3 Tone Contrast

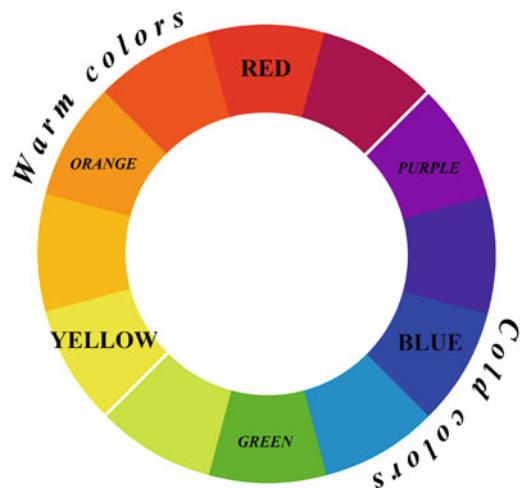
There are two forms of tonal contrast, global and local. Global tonal contrast measures the separation between the overall lightest and overall darkest areas of the entire image. Photographers such as Edward Weston and Ansel Adams have been influential in promoting the full use of the tonal range within the constraints put by the film and printing paper in fine art photography printing. A photograph with **high global contrast** displays black dark tones and white light tones. It evokes drama and dynamism. A photograph with **low global contrast** displays a compressed range of tones, with a limited range. It evokes softness and calm (Fig. 11.3).

Local tonal contrast is the separation in tone between adjacent areas. **High local contrast** lends to the perception of a high level of detail and depth, whereas an image with **low local contrast** appears flat and with less detail (Fig. 11.4). Local tone contrast evokes moods similar to global tone contrast, but note that the two are not necessarily linked: an image with high global contrast can have low local contrast: think of a gradient of light ranging from black to white.



Fig. 11.4 Low/high local tonal contrast. *Ocean, Redwood National Park, California, 2015. Mesquite Dunes, Death Valley National Park, California, 2005*

Fig. 11.5 RYB color wheel



11.2.4 Color

Hue, corresponding to the casual meaning of “color”, depends on the color’s dominant wavelength. Color theory establishes that three primary hues are sufficient to generate all of them by mixing. However, there is no single, unified system for organizing hues, and the principles below apply to each of them. The traditional arts are dominated by the subtractive RYB color model, which was the first color model, originating in the early seventeenth century [6]. It is based on the three primary hues of red, yellow, and blue. Mixing them leads to the secondary hues of green, purple, and orange (Fig. 11.5). Each hue evokes many meanings, often subconsciously. The experience of our environment has led to some being universally shared, while others are symbolic and have been culturally acquired [15]. For instance, red could mean passion and vitality; yellow happiness and light; blue distance and coldness. Red, yellow, and orange are **warm colors**. Green, blue, and purple are **cold colors**. The first group increases tension, while the second releases it. Goethe first remarked



Fig. 11.6 Low/high saturation. *Badlands National Park, South Dakota, 2013.* *Grand Prismatic Spring, Yellowstone National Park, Wyoming, 2017*

that even if their luminance is equal, the first group is perceived as brighter, whereas the second group is perceived as darker.

Saturation measures how intense a color is, ranging from pure hue to grey. Color saturation plays a role similar to tone contrast and correlates with more emotion and expression. **Saturated colors** evoke drama, dynamism, and vulgarity, whereas **muted colors** evoke softness, calm, and elegance (Fig. 11.6). The same color hue could display a range of vastly different connotations depending on its saturation and luminance.

11.2.5 *Color Contrast*

The richness of color information is amplified when colors are considered through their relationship to each other rather than on their own. The perceptual basis for the relationships is the phenomenon of afterimage, where after staring at a color patch, looking at a blank area produces the sensation of the complementary color. The right mix of two complementary colors forms a neutral color, that is a shade of grey. Purple, green, and orange are respectively the complementary colors for yellow, red, and blue. They sit at opposite poles of the color wheel (Fig. 11.5). **Opposite colors** create drama and dynamism, with the most important contrast being between cold colors and warm colors. On the other hand, the effect produced by **neighbor colors** that sit close together on the color wheel is softness and calm (Fig. 11.7).

11.3 Secondary Visual Elements

Tones and colors are how we primarily perceive light or its absence in a photograph. From their interaction and contrast, discrete elements of geometric nature emerge, that represent the first level of abstraction of the photograph, bringing to it order and structure. Those elements are the basic building blocks of all visual



Fig. 11.7 Neighbor colors (yellow, green, green-blue)/opposite colors (orange, blue). *Cades Cove, Great Smoky Mountains National Park, Tennessee, 1998. Delicate Arch, Arches National Park, Utah, 1996*

communication, not just photography. The basic visual unit is the point or dot, but due to the continuous tone nature of photographs, dots are rarely seen in them. A set of dots so close together that they cannot be distinguished forms a line. The abstraction represented by secondary visual elements make them largely invariant to limited alterations of the primary visual elements that may happen in automated or intentional image processing.

11.3.1 Lines

There are two types of lines. Actual lines are either linear elements such as branches or, more frequently, edges between two areas of different tones or colors. Virtual lines can be defined by a direction of gaze or the direction of motion of a subject in the image. They also emerge by linking several small elements together. Because of the human need to remain upright during waking hours, the sense of stability is one of the most fundamental kinesthetic perceptions on which we depend for well-being. Vertical and horizontal axes are the basis for balance in a different way. **Horizontal lines** are at rest and equilibrium, communicating calmness and stability. **Vertical lines** also communicate stability, but in a more temporary way as things that stand vertically may eventually topple. They evoke strength and order. On the other hand, **diagonal lines** conjure instability and dynamism. **Curves** in lines add a touch of grace, delicacy, warmth, and are linked with an organic feeling (Fig. 11.8).

11.3.2 Shapes

Shapes are formed by a set of lines closing upon themselves. Small shapes act like dots, whereas the character of larger shapes is better defined. Although there is an endless number of irregular shapes, especially in nature, following our tendency to



Fig. 11.8 Horizontal, vertical, diagonal, curved lines. *Haleakala National Park, Hawaii, 2011. Voyageurs National Park, Minnesota, 2001. Bartlett Cove, Glacier Bay National Park, Alaska, 2001. Coyote Buttes South, Vermilion Cliffs National Monument, Arizona, 2019*

favor order and structure, we tend to reduce this chaos to a few basic geometric shapes typically found in human-made environments. Combinations and variations of those basic shapes make it possible to derive all other shapes. **Rectangles** are linked with formality, straightness, and balance; **triangles** with energy, conflict, and tension; **circles** with completeness, warmth, and spiritual connection (Fig. 11.9).

11.3.3 *Textures*

Textures emerge from the repetitive juxtaposition of lines or shapes. They are related to the material structure of surfaces, and as such serve as a visual equivalent to the sense of touch. Textures can be emphasized or minimized by camera angle and light angle. Low local contrast yields the perception of smoothness. High local contrast and detail level are associated with rough, hard textures. Sizes being equal, a more **textured area** attracts more attention with its dynamism and tension, whereas a **smooth area** is calming (Fig. 11.10).

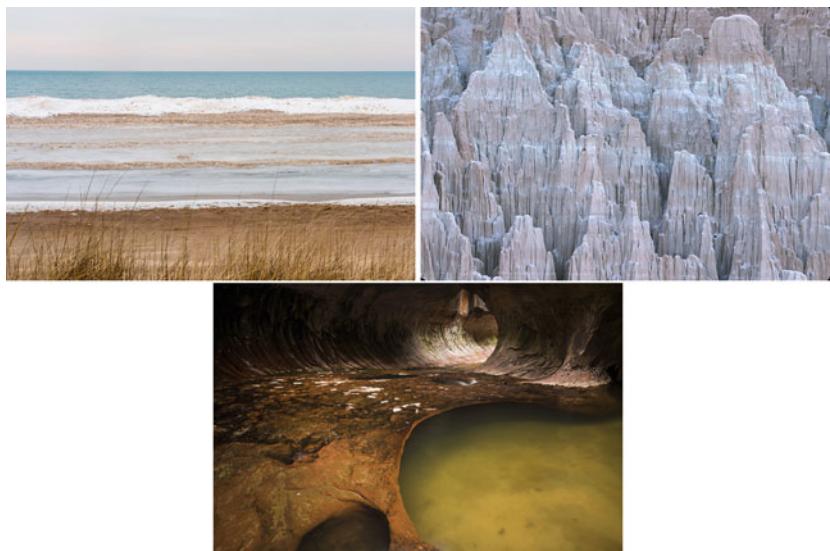


Fig. 11.9 Rectangles, triangles and circles. *Paul Douglas Trail, Indiana Dunes National Park, Indiana, 2019. Cathedral Gorge State Park. Nevada, 2003. The Subway, Zion National Park, Utah, 2017*



Fig. 11.10 Smooth/textured areas. *Ibex Sand Dunes, Death Valley National Park, California, 2015. Devil's golf course. Death Valley National Park, California, 2000*

11.4 Representing Three-Dimensional Space

Confronted with the need to describe the three-dimensional world on a two-dimensional surface, different cultures have produced different solutions, such as a flat calligraphic space in ancient Egypt and orthographic projection in traditional Japanese art. Photography owes its core visual language to developments that started in Italy during the Renaissance when the principles of linear perspective were discovered. Although the visual elements of a photograph are two-dimensional markings, reversing the principles of perspective makes it possible to infer three-dimensional properties.



Fig. 11.11 Front lighting, back lighting. *Gorman Hills, California, 2008. Joshua Tree National Park, California, 2013*

11.4.1 Light

Regardless of the subject matter, there are two things that all photographs depict: light and three-dimensional space. Light eventually results in the primary elements of tones and colors. Because certain light directions result in certain images, light can be identified from image elements. **Front lighting** (when the light source is behind the camera, therefore hitting the front of the subject) emphasizes colors but minimizes shadows, tonal contrast, and perception of shape. **Back lighting** creates silhouettes, where color is de-emphasized but shadows, tonal contrast, and perception of shape are maximized (Fig. 11.11). **Side lighting** combines attributes of both. **Diffuse lighting** happens when light comes from an extended area such as the overcast sky rather than from a specific location or direction.

11.4.2 Depth and Dimension

Three-dimensional space is the underlying structure of everything seen in the world. The three-dimensional space encompassed in a photograph can be either shallow, the extreme case being a plane parallel to the picture plane, or it can be deep and receding, extending from close foreground to the horizon. **Shallow space** creates a sense of opacity and calm, whereas **deep space** appears more transparent and dynamic (Fig. 11.12).

Humans perceive dimension directly through parallax, either with binocular vision or motion. In all two-dimensional representations of reality, dimension is only implied. The optical lens shares some of the properties of the eye, among them a simulation of dimension through perspective. Perspective clues in images make it possible to infer a sense of depth from a single image. A picture can describe deep space, yet present few perspective cues, and vice-versa. While the formal elements in the previous sections can be readily identified with simple feature detection, perspective clues require a more complex parsing. They include



Fig. 11.12 Shallow/deep space. *Long Canyon, Grand Staircase Escalante National Monument, Utah, 2020.* *White Pocket, Vermilion Cliffs National Monument, Arizona, 2019*

atmosphere, converging lines, layers, and scale effects. Nearer objects of similar size appear larger in the image. On the other hand, objects in photographs are large only relative to something small within the frame.

11.4.3 Volume

Gradual variations in tones of an imaged object can result from either gradual variations in reflectance of the original object, or interaction of light with shape. Our brain has learned to favor the second scenario because we assume by default that objects lie in three-dimensional space and that they have varying sizes and shapes. When we see two-dimensional shapes with gradual progression from light to shadow, we read them as **volumes** with shadowing. Moreover, we have also learned that light generally comes from above and therefore we interpret the patterns of light and shadow as indicative of the height of an object: a lighter top and darker bottom indicate a round object, whereas the opposite pattern indicates a hollow.

11.5 Photography-Specific Elements

With the exception of black and white tonality, the elements described in the previous sections have been inherited from older art forms such as drawing and painting. However, photography brings specific characteristics that determine how a camera transforms the world into a photograph [10]. While it is possible that AI-generated images will eventually come to be accepted as *photographs*, by the current definition of the term, they are *not* photographs because they are not made through a camera with the implied attributes elaborated on in this section.

11.5.1 Vantage Point

It is the invention of photographic emulsions, rather than the camera, that distinguished photography from previous art forms. Long before the invention of photography, the camera obscura was used as an aid in creating images, but the artist had the eventual license to place elements as desired. By contrast, photography deals with the actual [13]. What the emulsion records is the projection of the three-dimensional world into a flat imaging surface through a single, definite vantage point. That transformation creates new relationships between lines and shapes that did not exist in the world, relationships that are changed by any change in the vantage point. The choice of a vantage point also determines what in the scene becomes the **foreground** as opposed to the **background**.

Because the camera has to be placed somewhere, by necessity photographs provide us with a larger variety of vantage points over the world than any art form before—the work of Alexander Rodchenko is characterized by unusual vantage points. The whole range of camera heights from bird's view (overhead) to bug's view (low) translates to a whole range of camera angles. A straight, **eye-level angle** suggests directness and honesty, whereas an **oblique angle** creates tension, especially if it causes an extreme foreshortening distortion. **High vantage points** evoke omnipotence and detachment; **low vantage points** evoke vulnerability and awe (Fig. 11.13).

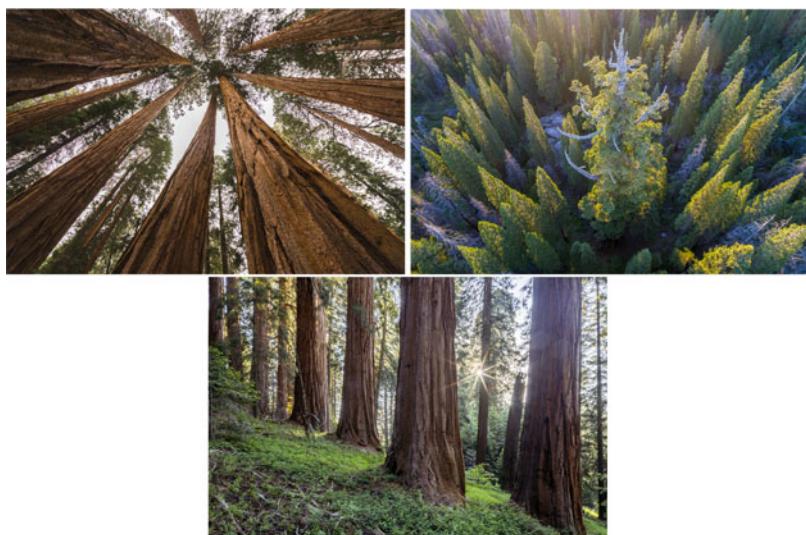


Fig. 11.13 Low/High vantage point with oblique angle, eye-level angle. *Sequoia National Park, California, 2015. Boole Tree, Giant Sequoia National Monument, California, 2020. McIntyre Grove, Giant Sequoia National Monument, California, 2020*

The use of different lenses opens up new avenues of expression as they allow the distance from which the photograph is made to be varied. This alters the relationship of the photographer to their subject, for example from a hidden voyeur with a telephoto lens to a participant with a wide-angle lens. Wide-angle lenses dramatically magnify perspective effects, increasing apparent distances and making foreground objects much larger than background objects. That distortion of space, noted in the work of Bill Brandt, is so unnatural that it had seldom been used in the visual arts before photography. Telephoto lenses generate the opposite effect, flattening space and making objects look closer than they are with a projection akin to orthography. Images that offer a natural visual experience are calmer and more harmonious than distorted images that deviate from realism.

11.5.2 *Frame*

Unlike the world, a photograph is bounded by edges. Photography has inherited the rectangular frame from other visual arts such as painting. However, the role of the frame in photography is drastically different from other visual arts. The painter fills an initially empty frame with visual elements, making something more complex than what he started with, a synthetic process. The photographer works in an opposite direction, simplifying and putting order into a chaotic world, an analytical process. The photograph is selected, not conceived [13]. The imposition of a frame determines what is in the image and what is not. Before photography, visual elements were organized within the frame as if it created a window through which the viewer looked out into the world. The randomness that comes with pointing a camera at the world broke that prescription. Photographers discovered that the world could be arbitrarily truncated, with detached elements or shapes partly created by the frame forming successful compositions. This new visual vocabulary in turn influenced painters such as Edgar Degas.

The edges of the frame are prominent lines in the photograph and as such are subject to the same observations we made for its line elements within. The **shape of the frame** has often been traditionally determined by the film format, from the neutrality of the square frame to elongated panoramic formats that by themselves add a sense of complexity. A horizontal frame causes the viewer's eyes to move horizontally, whereas a vertical frame favors a vertical scanning of the image. Edges create relationships between the elements within the frame and also with the frame itself. Some pictures are more actively structured by the frame than others. Their self-contained structure works inward from the **active frame** that determines the relationship of their formal elements. For others, typically portraits, the structure works outwards and the role of the **passive frame** is simply that the picture has to end somewhere (Fig. 11.14).

Humans have a basic need for stability while interacting with their environment. As the lack of stability is a disorienting factor, photographs tilted sideways are an exception, and in general, the frame is aligned with the vertical and horizontal



Fig. 11.14 Passive/active frame. *Great Sand Dunes National Park, Colorado, 2009. Meadow Overlook, Shenandoah National Park, Virginia, 1998*

axes. The vertical axis in the middle of the frame (the “felt axis”) is the primary reference for balance and therefore is what our eyes tend to scan first [4]. Although a rectangular frame appears geometrically symmetrical, it is not so perceptually. As noticed before, humans assume light comes from the top. They also assume the ground is at the bottom, a part of the frame scanned by our eyes second. Elements placed at the bottom of the frame tend to be at rest, while elements at the top of the frame call attention to their instability. In addition, the Western mind generally is trained to read from left to right. While we are reading a picture, our eyes tend therefore to start from the lower left quadrant. An object that is placed in the right or top half of the image will carry greater weight than an identical object that is placed in the bottom or left half.

11.5.3 Time and Movement

Unlike in other arts, photographs depict only that time during which they were made, hence are fundamentally unable to represent history. A photograph transforms a fluid world into a static representation in a way determined by the duration of the exposure. Photographs are not instantaneous, but describe a section of time. If nothing is moving, the photograph records **still time**, which is interchangeable. When there is motion involved and the exposure is short enough that the change between its beginning and end is minimal, the photograph freezes a moment in time. Using electronic flash (as pioneered by Harold Edgerton) freezes the fastest of movements. It was not possible to immobilize those fast events before the invention of photography. Capturing a **frozen instant** (Henri Cartier-Bresson’s “decisive moment”, a visual climax) is central to photographic storytelling.

A long exposure during which the subject or the camera moves records a **motion blur**. The simplest and most frequent case occurs when the camera is fixed, resulting in the moving parts blurred against sharply rendered stationary parts of the scene. Depending on the length of the exposure, the direction of motion could be seen, resulting in a dynamic effect, or the blur could be uniform like in long exposure



Fig. 11.15 Blurred/frozen water motion. *Berry Creek Falls, Big Basin Redwoods State Park, California, 2003. Nevada Falls, Yosemite National Park, California, 2010*

of water, resulting in a soothing effect (Fig. 11.15). The camera can also track the subject's motion, either from a stationary or moving position, resulting in a blurred background in the direction of the camera's motion. Motion blur is an addition to the visual vocabulary that can be attributed solely to photography. Once photographers invented these ways of conveying motion, viewers learned those conventions so that they could understand the photographs. Subsequently motion blur has been appropriated by other graphic arts and is now universally recognized as signifying motion and change.

11.5.4 Focus and Depth of Field

Photographs are imaged by lenses. Lenses can be focused on only one plane at a time. Unless the photographer uses an unusual camera system that offers flexible positioning between the lens and the imaging plane, the plane of focus is a plane parallel to the image, and therefore characterized by its distance to the camera. In theory, objects in front or behind that focus distance will not be imaged sharply but will be blurred, as points are rendered into circles. This **focus blur** is unique to lens-based imaging.

In practice, because the human eye has limited resolving power, a circle up to a certain size is perceived as a point, and therefore the corresponding object appears sharp. The area in the world for which this happens is the depth-of-field, and its extent increases with the focusing distance and the field of view of the lens among other factors. Two equipment choices that lead to a diminished depth of field are wide aperture lenses and large image capture areas, both of them rarely used by casual photographers. A **shallow depth of field** concentrates the viewer's attention on the area in selective focus, creating a hierarchical emphasis in the photograph. Because it has fewer blurred areas, a photograph with a **large depth of field** displays more detail and appears more realistic (Fig. 11.16). The presence of blur softens an image and can make it more mysterious or dreamy. Sometimes, independently from the depth of field, filters or specialized lenses are used to render the entire image



Fig. 11.16 Shallow/large depth of field. *Antelope Valley California Poppy Reserve, California, 1994. Carrizo Plain National Monument, California, 2017*

with less sharpness, which introduces the distinction between overall sharp focus and soft focus. Sharply rendered images have a tense edginess to them, whereas soft images are perceived as more harmonious.

11.6 Composition

If in the visual language of photography, the words are visual elements such as tone, line, and frame, then the syntax is composition, which refers to how a photograph is arranged or constructed. Visual element types are limited in number, and it is possible to identify them as we have done. On the other hand, visual composition techniques are more challenging to categorize and enumerate. However, one useful way to classify them is to use a basic principle rooted in the psychology of perception, the polarity between harmony and contrast that we have already encountered multiple times when examining individual visual elements.

11.6.1 Harmony and Contrast

The human mind naturally seeks a state of ease, resolution, and calm. Maximizing harmony is an approachable way to creating satisfyingly calm images, but it can lend to a lack of stimulation, like in the Buddhist ideal of nirvana where sensations are extinguished. On the other hand, our attention is activated and stimulated by tension and contrast. Without the contrast of tone, an image is nothing but a uniform patch of grey, where there is nothing to see. Gestalt psychologists define two opposite visual states, “sharpening” or “stress” as an increase in exaggeration, and “leveling” as a weakening of a peculiarity of a pattern [7]. Those correspond to the concepts of contrast and harmony and are the basis for the basic polarities of visual composition: tension and calm. We have seen how the attributes of visual elements align along this polarity. In addition, a composition can contrast elements with polar attributes, for

example dark areas and light areas. The contrast in tone is the most important formal component of photographs, but beyond that, all the visual components could be contrasted: colors, textures, shapes - with irregular shapes attracting more attention than simple shapes.

11.6.2 Balance and Instability

As discussed before, because of the human need for upright stability, vertical-horizontal axes are the basis for our interaction with our environment. In visual terms, the need for balance imposes a vertical axis in the middle of the frame. Centering a subject creates a classical balance. When the important elements on the left side do not have a counterbalance on the right side, the image can feel tipped and unsettling. Balance, like harmony, is a leveling technique, while instability is a sharpening technique.

The simplest way to achieve perfect balance is bilateral symmetry, but this could be perceived as too much leveling. To keep a measure of tension, balance can be achieved asymmetrically by involving visual weight and position. To that effect, visualize a balanced scale with weights on both sides. The balance can be kept by count if on a side we replace one weight with two smaller weights of the same mass. A heavier weight placed closer to the scale's center will also balance a smaller weight placed further. The visual weight of an element may depend on its size or tonality: a large light area can have as much visual weight as a small dark one. The use of position for asymmetrical visual balance is illustrated by the "rule of thirds". The frame is divided into two sets of thirds, one horizontal and one vertical, forming a grid of nine rectangles, and the subject is placed at one of the grid intersections, where it is balanced by more space on the other sides (Fig. 11.17).

11.6.3 Repetition, Irregularity, Unity, Variety

Visual elements that bear a high degree of resemblance naturally harmonize with each other. Their repetition creates connections and a cohesive force holding a composition together. Uniform repeated elements could lead to a lack of variety. In the same spirit as dynamic balance, they could differ in some aspects while following a logical order, such as sequential progression or alternation, which create a visual rhythm through a regular and predictable plan (Fig. 11.18).

The dynamic polar opposite composition technique is irregularity, characterized by a loose connection between visual elements, randomness, and a lack of planned order. A composition with elements that retain a highly individual character presents variety, which contributes to interest. However, without some unifying elements, that variety could be confusing. A photograph is unified if all of its parts appear to contain a portion of the whole within them, or if it would break apart if one of its



Fig. 11.17 Symmetry, asymmetrical balance by position (rule of thirds), count and visual weight. *Goat Island Mountain, Mount Rainier National Park, Washington, 2011. Kabetogama Lake, Voyageurs National Park, Minnesota, 2001. Courthouse Wash, Arches National Park, Utah, 2013. Cinder cone, Lassen Volcanic National Park, California, 2000*



Fig. 11.18 Repetition/progression. *Acadia National Park, Maine, 1997. North Cascades National Park, Washington, 2010*

parts is removed. Its diverse elements mesh so well that they form a picture that is viewed as a single thing, greater than the sum of its parts.

11.6.4 Simplicity, Complexity, Emphasis

In a totally neutral composition, all the elements carry the same weight, and none stand out, like a background without a subject, leading to a calm impression. Emphasis is the addition of a center of interest, also called a focal point, which



Fig. 11.19 Single focal point/complex composition. *Ahwahnee Meadow, Yosemite National Park, California, 2007. Green Mountains, Vermont, 1997*

creates tension against the sameness of the background. A photograph containing a solitary focal point has a simple and direct composition. More tension is yet created when two or more focal points are juxtaposed, creating a complex and visually intricate composition. Simplicity and complexity are closely related to the notions of economy and intricacy measured by the number of visual elements present (Fig. 11.19). Those dual approaches are exemplified by photographers Michael Kenna and Alex Webb.

Photographers have numerous formal tools at their disposal to create emphasis: contrast—typically of sharpened visual features standing out against leveled features, arrangements of lines and shapes that lead the eye to a specific place, or placement in foreground areas. However, emphasis is often created by subject matter such as a life form in an otherwise inanimate landscape. Regardless of size, the human figure usually attracts more attention than anything else in a picture. Recognizing subject matter requires moving beyond a purely formal analysis.

11.7 Beyond Form

This chapter has focused on the abstract level of formal description as the aspect of photographs most conducive to automated approaches. However, photographs are read by humans at several other different levels, that involve respectively the world being photographed, the photographer, and society. This section briefly points out how a complete reading of photographs requires going beyond form and mobilizing a varied corpus of knowledge.

11.7.1 Form and Contents

While it is possible to identify most formal elements of photographs without reference to the things being depicted, others are tied to subject matter. Take

for instance the figure/ground relationship, which is maybe the most fundamental perceptual concept [15] and is present in all photographs. Formally, the figure is the positive, object-occupying space, whereas the ground is the negative space surrounding the objects. The relationship can be established with the help of visual elements, such as contrast, color, size, position, or focus. But humans generally establish it by identifying what the objects represent in the real world.

Photographs are inextricably linked to reality. The identification of what a photograph represents is of primary importance to our perception of it. Viewers make the assumption that every photograph represents a piece of the world, and know how to react to it only when they know what it is [12]. Some photographs derive their force from direct tie to traumatic events (the “Punctum” [2]), when the contents overwhelm all other aspects of the image, for example in Nick Ut’s *Napalm Girl* (1972). On the other hand, when looking at a still-life photograph of a common object by a photographer pursuing perception for perception’s sake, the form may dominate. Most photographs lie in a continuum between those two extremes, but rarely transcend entirely the contents.

11.7.2 *Contents and Subject*

A photograph needs to be of something existing in the world, its contents, or subject matter. William Henry Fox Talbot, one of the inventors of photography, envisioned it as an impersonal way for nature to record itself accurately, but it turned out that different people took different pictures of the same subject matter [12]. A photographer brings his unique, subjective, point of view (how unique determines the originality of the work), which leads to the photograph’s subject. The subject matter is what the photograph is “of”, its explicit contents, whereas the subject is what it is “about”, its implicit meaning. Photographs fall in a continuum between being almost entirely about the subject matter (such as scientific photography) or about the subject.

In literature, metaphors convey a different meaning than the literal, creating illuminating associations. In the language of photography, a visual metaphor is created when the photographer uses creative choices to portray a thing as something other than what it is. This allows photography to communicate abstract ideas, feelings, or intangible aspects of reality beyond the physical surfaces that it can only directly depict. Alfred Stieglitz asserted that photographic meaning is metaphoric as illustrated by his *Equivalents* (1925–1934) series.

11.7.3 *The Photograph as Object*

There are aspects of a printed photograph that can be missed if it is viewed as a disembodied digital file. Some of the physical qualities of a print are reflected in

the formal characteristics of photographs: that the print is flat, has edges, displays specific tonal characteristics dependent on its materials, and is static [10]. What an art object is made of, its medium, including physical size and reproduction quality affects how it is perceived. Photography is one of the few art forms where works can be original and reproducible at the same time. The photographic print possesses an “aura” [3], which is the emotional response of viewers standing before an original work of art (as opposed to a photomechanical reproduction), a result of the object’s history and physical presence.

Photographs, by their nature of being a slice of space and time, are out of context. Knowledge of the original context in which the photograph was made can be crucial in interpreting the photograph. As an object itself, a photograph lives its own life in the world. The external context in which it is seen affects the way viewers read it. The same photograph used as a magazine advertisement, a photojournalistic story, or a museum exhibit elicits different meanings that can easily override its contents [5].

Besides existing materially in the world, photographs are also cultural objects. A small fraction of them endures as artworks part of humanity’s shared cultural and artistic history. But eventually, all photographs become, with time, documentary [9]. As the image moves from the instantaneous to the moment, even vernacular works change from being substitutes for reality to the way we represent things to ourselves.

References

1. Barrett, T.: Criticizing Photographs: An Introduction to Understanding Images. McGraw-Hill, New York (2006)
2. Barthes, R.: *La Chambre claire: Note sur la photographie*. Gallimard, Paris (1980)
3. Benjamin, W.: The Work of Art in the Age of Mechanical Reproduction (1935). Translated by J. A. Underwood. Penguin Books, London (2008)
4. Dondis, D.: A Primer of Visual Literacy. MIT Press, Cambridge (1973)
5. Freund, G.: *Photography and Society*. David Godine, Boston (1980)
6. Goethe, J.W.v.: Theory of Colors (1810). Translated by Charles Lock Eastlake. MIT Press, Cambridge (1982)
7. Koffka, K.: Principles of Gestalt Psychology. Harcourt, Brace, and World, New York (1935)
8. Peirce, C.: Collected Papers of Charles Sanders Peirce, Volumes I and II: Principles of Philosophy and Elements of Logic. Harvard University Press, Cambridge (1932)
9. Scott, C.: The Spoken Image. Photography and Language. Reaktion Books, London (1999)
10. Shore, S.: The Nature of Photographs. The Johns Hopkins University Press, Baltimore (1998)
11. Singh, R.: River of Colour: The India of Raghubir Singh. Phaidon Press, London (1998)
12. Sontag, S.: On Photography. Farrar, Strauss, and Giroud, New York (1978)
13. Szarkowski, J.: The Photographer’s Eye. The Museum of Modern Art, New York (1966)
14. Szarkowski, J.: William Eggleston’s Guide. The Museum of Modern Art, New York (1976)
15. Zakia, R.: Perception and Imaging. Focal Press, Oxford (2007)

Chapter 12

Breathing with Robots: Notating Performer Strategy, Alongside Choreographer Intent and Audience Observation, in Breath-Driven Robotic Dance Performance



Kate Ladenheim, Amy LaViers, and Catherine Maguire

Abstract This chapter presents an analysis of the choreography for *Babyface*, a dance solo and interactive installation comprising a pair of breath-activated robots. A discussion of breath from somatic and choreographic perspectives is offered alongside new symbols for the description and notation of breath. The paper also highlights an important triptych in the creation of meaningful movement—the choreographic vision, the performer’s strategy, and what can be observed by the audience—which is heightened due to the breath-based control mechanism for the mechanical device. These elements illustrate how visual aesthetic, style, and emotion work together in any salient moment of human-robot interaction, and the analysis presented proposes a systematic way of accounting for breath across these elements. Moreover, in taking a myopic and enthusiastic interest in a particular work of choreography, the chapter demonstrates how practical ends are enabled by this curiosity-driven research at the intersection of robotics and the performing arts.

12.1 Introduction

Of the many bodily modalities humans use to connect with one another—e.g., vocalization, locomotion, and focus—breath may be the most visceral and intimate—and

K. Ladenheim

School of Theatre, Dance, and Performance Studies, University of Maryland, College Park, MD, USA

e-mail: klad@umd.edu

A. LaViers (✉)

Robotics, Automation, and Dance (RAD) Lab, Philadelphia, PA, USA

e-mail: amy@theradlab.xyz

C. Maguire

WholeMovement, Palmyra, VA, USA

one of the least utilized in human-robot interaction. One challenge in utilizing and understanding breath is detection: traditional detection systems utilize inconvenient masks worn over the mouths of wearers or indirectly detect rate from heartbeat monitors. Moreover, facial expression, the language content of vocalizations, and the gross bodily motion of major joints are often more apparent—especially to a lay observer—in observations of human motion than subtle shifts in the spine or changes in breath rate.

As to the *end goals* of the work presented in this chapter, prior work has investigated breath in the context of human-robot interaction. Robots have been used to explore facilitating deep breathing in human counter parts for stress management [19]. Breath detection has also been investigated to improve the quality of care-giving robots at the bedside of patients [21]. Moreover, it has been established that, like gaze, breath detection may be important for social cue design in collaborative robots for manufacturing [24]. While requirements for wearable sensors make breath detection a challenge in many HRI scenarios, progress toward contactless breath detection has been made [25].

As to the *process* of the analysis presented in this chapter, using curiosity-driven research at the intersection of robotics and dance also has precedent in prior work [8]. The role of aesthetics in embodied meaning-making has been presented as central to the investigations of robot design for HRI [6]. Exploring the limits of the Turing formulation of computation in a participatory dance performance [15] coordinated with the development of a model for robot expressiveness [14]. User studies were run in parallel to an onstage performance [3] of the work “Time to Compile” [4].

Biomechanics, dance instruction, and human-computer interaction also offer inroads to understanding the role of breath and measurement in technology. Conditions called breathing pattern disorders have been found to inhibit the functional success of movement [1]. In teaching dance, many instructors encourage vivid visualization of breath for students [20]. Breath patterns have also been used for cues inside human-computer interaction research [2].

This chapter presents an analysis of a choreographic and engineered work called *Babyface* that developed movement expression of human performers alongside a novel machine to extend the action of these live performers (as well as audience-based participants) with mechanical actuators. The team of equally contributing authors, who all share interdisciplinary backgrounds in both dance and robotics, includes two primary creators of *Babyface* (Ladenheim and LaViers) as well as two experts in movement analysis and notation (LaViers and Maguire). Not being connected to the creation of *Babyface*, Maguire offers an outside perspective to the piece that Ladenheim and LaViers do not have, given their familiarity with the work. Moreover, as choreographer and performer, Ladenheim can offer insight into the makings of the onstage phenomena, on which our analysis centers.

To analyze the role of breath in *Babyface*, we suggest a new notational convention to identify distinct breath cycles in Sect. 12.2. The work *Babyface* is then described

in Sect. 12.3 and key moments from the work are highlighted for further examination. Section 12.4 will offer exercises for the reader to practice and experience these distinct cycles as well. We will then introduce a three-part analysis in Sect. 12.5 that includes a description of the choreographer’s vision and three motifs (a type of movement notation): one of the performer’s strategy, one of the resulting machine action, and one of the overall impression of performer and machine in context. This work’s immediate and future application to human-robot interaction more broadly will be discussed in Sect. 12.6. Finally, a concluding summary is offered in Sect. 12.7.

12.2 Breath

Breath is one of our most foundational movement patterns. As a foundational medium of expression, how we breathe will not only change our own experience but the observation of our experience by others. The pattern of inhale/exhale is both involuntary (occurs whether we put our awareness to it or not) and volitional (we can change the pattern, rhythm, duration, emphasis etc.). This pattern is also responsive to stress, exhaustion, excitement, and our reaction to the environment broadly. That is, the inhale illuminates our inner body and our exhale connects us to the outer world.

A somatic, experience-based view of the body emphasizes the three-dimensional nature of breath with irregular expansions in all three dimensions. This view often emphasizes the core as central to bodily expression and de-emphasizes the role of the highly deformable limbs. One such somatic practice, Laban/Bartenieff Movement Studies (LBMS), describes this through primary dualities of inner/outer and self/other: all that is me (inhale, inner, self) connecting to all that is not me (exhale, outer, other) [23]. The involuntary act of breathing becomes functional and can be understood as a state of “being” while the volitional breath invokes a state of “doing”. These ideas allow us to understand breath as a medium of our expressive self, or a medium of expression.

Breath allows us to experience the continuum of inner/outer in terms of innersphere (our internal 3D volume), kinesphere (the space around our body that we can reach without locomotion), coronasphere (the external space of the environment that our breath reaches [5]), and general space (the greater environment in which our actions translate to movement). We can more granularly define the macro shape change (the bridge between body and space) pattern of growing and shrinking to internal awareness of our 3D internal core space: lengthening and shortening (vertical space), widening and narrowing (horizontal space) and bulging and hollowing (sagittal space).

Moreover, the quality, depth and pattern of our breath reveals our inner state—think of “breathless”, “take my breath away”, “holding my breath”, “breathe life

into it”, as macro ideas. More granularly, as breath connects us to our ongoing sense of flow the expressive support of our intent becomes manifest. The primary sense of the ongoing pattern of release and control can lead to more explicit expression of inner intent through the Effort component of LBMS [10]. Think of a forceful exhale (strong weight effort), a wispy sigh (light weight effort), blowing out candles (direct space effort) shushing a room of people (indirect space effort), gasping with surprise (sudden time effort), lingering in a prana breath for relaxation (sustained time effort), holding our breath (bound flow effort) and releasing that held breath (free flow effort). Thus, the quality of the two part breath phrase and its ongoing pattern of release and control directly supports the expression of inner intention.

Shape is important in revealing the breath cycle—can we actually “observe” the breath? Or are we in fact observing the resultant shape changes in the body? So while the dualities of inner/outer and self/other are central to the experience of breath, we discovered that the idea of growing/shrinking, is central to the perception of breath. In other words, to perceive breath we need to observe the body’s changing form in relationship to the environment.

In LBMS breath is identified as a foundational movement pattern as mentioned above, but little has been done in terms of trying to notate the breath and the breath cycle. Hackney identifies breath as a pattern of body organization—using the symbol for the “body” (essentially a figure-eight shape) with a circle in the center [7]. In our application we identify breath as a primary body based cycle that, because it is both involuntary and voluntary, underlies all movement experience.

To notate choices in breath patterning, we utilize previously established symbols for breath [22]. These symbols use a circle with dotted lines inside the circle and dotted lines outside the circle to notate “breath.” The symbol can then be separated with just the circle and dotted lines inside to represent inhale and just the circle and dotted lines outside to represent exhale. Using these symbols together inside a vertical motif [9], we can indicate the temporal relationship of the parts of the breath cycle to each other, thus creating a notational abstraction [18] that provides a perceptual capture on the breath cycle itself. Thus, Fig. 12.1 shows symbols for breath, inhale, and exhale, which can be combined and stretched to show temporal sequencing and duration for different styles of breath.

The next section introduces the work of *Babyface* and identifies breath patterns for further analysis with this symbolic taxonomy.



Fig. 12.1 Symbols for breath [18, 22]. Left: breath; center: inhale; and right: exhale. Stretching in the vertical dimension inside a vertical motif indicates longer duration. Ordering in sequence provides information about breath over several cycles

12.3 Breath Patterns in *Babyface*

The artistic work *Babyface* [11–13, 16, 17], shown in Fig. 12.2, is a dance and robotics installation that reflects on feminine stereotypes in media and robotics design. It uses breath as a somatic channel between body and machine, as well as a metaphor for the depth of scrutiny and efforts towards control that are levied at feminine bodies, whether they be human or artificial. The following exercise brings a moment of the tone, mood, and content of this performance to these pages. It is best performed in front of a mirror:

Start by taking a deep breath in, and a deep breath out. You look wonderful. Smile! Place your hands on your hips, and spread your feet just wider than your shoulders. Yes, that's it! Don't move your feet or your hands, but push your right hip out to the side. Look up, and to the right. Breathe in. Breathe out.

Look at yourself in the mirror, at the shape of your body as you breathe in, and breathe out. Do you feel beautiful?

Now, place both of your hands behind your head. Elbows out to the side. You're becoming the best version of yourself, it's magical.

Now move your hip to the left as you breathe in, and to the right as you breathe out. That's very good...do you feel good now? Do it again - breathe in and push your hip to the left, and breathe out as you push your hip to the right. Keep your hands behind your head. Keep going, finding a regular rhythm. Try a little faster. Can your breath keep up with your body? Are you staying in control?

Do you still feel beautiful?

Take a moment to reflect on this exercise, and the combination of breathing, moving, and self-inspecting that the exercise offered. How did it change or reframe your usual experience of breath?

Breathing does not have an inherent, singular meaning, and yet this bodily action communicates something. That something changes, and is dependent on contextual markers: the environment, cultural training, social cues, and more. These factors can be manipulated by artists, the way we just did with the above exercise, which is adapted from the performance. Our stylistic choices and aesthetic markers (such as our enthusiastic, inspirational tone; short and to the point sentence structure, and leading questions about beauty and self-improvement) created associations that we designed through tone and physical movement requests. Dance artists are especially skilled at crafting meaning from motions that, like breathing, do not have one static meaning or specific emotional resonance.

Babyface offers audiences two modes of interaction that try to create emotional resonance through breath. The work invites audience members to activate a pair of large scale robotic wings with their breath via a wearable pressure sensor that detects the expansion and contraction of the torso. The wings augment each audience members' body, simultaneously serving as an impressive, even delightful spectacle and a physical metaphor for the outsize expectations placed on feminine bodies via technological design and social pressure.

The work also includes a performance from a cyborg character who activates wearable wings with her own breath and motion. Like the audience member, the winged performer is an impressive spectacle, but she is inescapably framed by the

machine she wears and the retro-futuristic performance design. The delight of being regarded (for the performer as a character) and the enjoyment of a cute, machinated woman (for the audience) contrasts sharply with the stereotypical rigidity of hyper-feminized performance. This is the core tension of this work, and these oppositions are navigated and mediated through breath.

For an audience member, the payoff of moving the machine with their breath requires an internal, almost singular focus on their breathing; audiences must execute a high level of control over the motion of their body, and furthermore, over a motion that is usually unconscious and automatic. That is to say, they have to exaggerate the motions of their breath, and conform their motions to the style that is instructed. This is integral to the choreographic intent of this interaction: breath here is not only the driver for the machine's actuation, but also a metaphor for how women must often be hyper-aware of the way their bodies perform at all times. This larger societal pattern has a profound emotional impact on women writ large. However, not all women feel this pressure in the same way, or manifest the same emotional response. Some women take genuine pride and enjoyment in the attention, while others find it alienating; others still barely notice it. This plays out in reactions to the installation as well as in society.

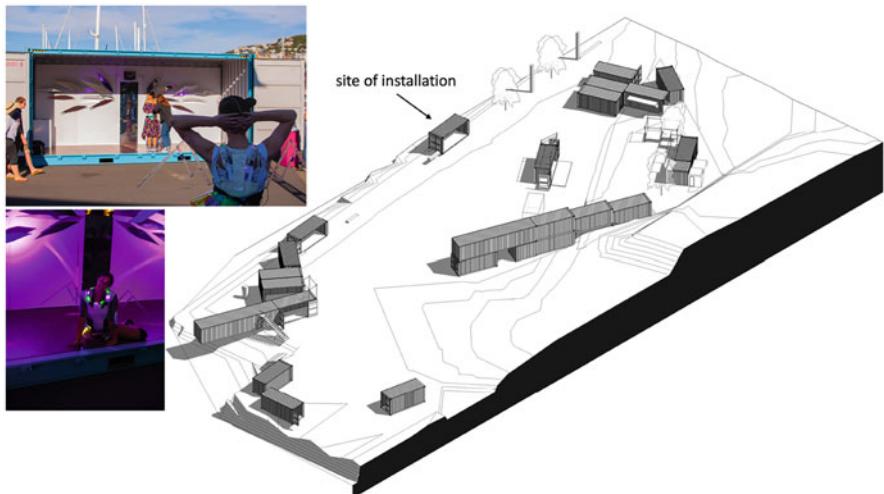


Fig. 12.2 A diagram of the public installation in Wellington, NZ. The installation was part of the Performance Arcade, which houses public artworks in multiple containers temporarily deposited along the waterfront of the city (drawn above). The *Babyface* installation and performances (pictured at day and night at left) were accessible to show attendees and passersby on the waterfront promenade. Any audience member that wanted could stop at the site and try on the breath sensor (being attached to a person in a floral dress in the daytime image above) in order to activate the wall-mounted robotic elements with the expansion and contraction of their core due to breath. Images: Colin Edson and the Performance Arcade

The performer also must be especially attentive to her breathing; indeed, it is the way that she synchronizes the machine's expression and the expression of her body in line with the overall choreographic vision of the work. Even so, breath is not always—or even usually—the most salient visual component for the audience. Simply put, there is a lot more going on. There is a structured choreographic vocabulary; a sound score complete with voiceover; the motion of the machine on the performer as well as mounted on the wall; lights and costuming. All of these components together build an aesthetic world for *Babyface*, and breath is but one part of it—one that supports motion fluency and expressive aims, but stops short of being the primary visual impression. Rather, it augments. As with the participatory experience, breath as a metaphor for scrutiny is materially important to the work's overall message. When combined with stylized choreography, it takes on aesthetic qualities that support an emotional message.

To this end, *Babyface* employs several breath patterns, and we focus on five in this chapter:

- **Breath Pattern 1** (*longer inhale, relative to a shorter exhale*): This breath usually supports preparation for motion in performance; the long inhale is intended to build anticipation or excitement, or to show that the performer is reaching or yearning for something.
- **Breath Pattern 2** (*shorter inhale, relative to a longer exhale*): This breath is intended to reveal a malfunction in cyborg character's presentation of self.
- **Breath Pattern 3** (*even, exhale followed by an inhale of similar duration*): This pattern involves a holding and repetition that is intended as a cutesy or even campy rhythmic punctuation.
- **Breath Pattern 4** (*longer exhale, relative to a shorter inhale*): A shrinking of the body and falling of the torso that hopes to communicate shame, disappointment, and/or exhaustion.
- **Breath Pattern 5** (*even, inhale followed by an exhale of similar duration*): Intended as an opportunity to regain control and focus; requires concentration and synchronization to perform.

Thus, by analyzing *Babyface* in moments where the breath cycle was an important strategy for the performer to engage in in order to express the choreography, we identify six breath cycles (the five identified above as well as the sixth enumeration in the same pattern). The even inhale/exhale, the even exhale/inhale, the long inhale/short exhale, the short exhale/long inhale, the long exhale/short inhale and the short inhale/long exhale. Each cycle is both functional and expressive in the context of the choreography.

The intended emotional impacts of the work (the choreographic vision), what the performer is attentive to and experiencing, and what the audience feels are not always in perfect alignment. The choreographer is responsible for how motion imparts intent; not just in the gross motions of the body and in the machine, but also in coaching stylistic delivery and responding to the aesthetics of the installation. The performer translates this choreographic vision, which requires very intimate awareness of how the machine works and how breath, body positioning, and acting can

make the choreographer's intent actualized. Skilled choreographers and performers can guide an audience towards an emotional response, but ultimately the audience's emotions are their own. In this way, style can be seen as a collaborative effort between choreographer and performer; aesthetics can be seen as a collaborative effort between choreographer, design team, materials, and the machine; and emotion can be seen as a collaboration with the audience rather than a direct, guaranteed manipulation. These distinct breathing states facilitate this collaborative exchange between body and machine, audience and creators, and contributes to an overall stylistic, aesthetic, and emotional picture.

The next section offers exercises for you, the reader, to try in your own body. This type of embodied investigation is essential to understanding concepts in movement studies [18]—such as those of breath introduced so far and used later in the paper for notational analysis of *Babyface*.

12.4 Embodied Exercises for Deeper Understanding

Breathe in, breathe out. Breathe in, breathe out. This ongoing cycle offers a wide breadth of choices: we can take a longer exhale than inhale, we can use sharp, fisted force with our exhale, or we can breathe softly, making the motion barely perceptible to others. Different environments and activities require each one; and each one changes our experience of every environment and activity we may choose. It is, quite literally, a way that we express ourselves in our environment (public or private), and it is kinesthetically bound with our emotions and our broader subjective experience, as the opening narration illustrates. That is, we may say that the ongoing cycle of movement present in every human body living today presents a large bandwidth of choice in style to fit within the aesthetic of a particular context or moment and to express our own point-of-view or emotion about it.

For your own experience of these cycles try the following movement sequences. As you approach exploring these breath cycles, begin by just settling into noticing your breath and tuning into an awareness that you can make volitional changes to how you experience your breath. Each exploration should begin by establishing a steady clapping rhythm—at any tempo you desire. Each exercise offers a more mechanical, directed inroad to the breath style, which may be practiced at differing tempos, as well as a more open-ended, image-driven experience, which, optionally, offers greater depth with additional choreography that the reader can explore as well.

1. Practice Breath Cycle In/out Even: After establishing your clapping rhythm try inhaling on a count of three claps and exhaling on a count of three claps. Allow yourself to find an even breath cycle.

- Imagine you are sitting reading a book that is on your lap. You want to stretch your back out and then settle back into reading. Stand up and stretch on the inhale and settle back down to sit on the exhale.

- To heighten your experience, as you inhale and stand up allow your chest to rise and advance with your head tilted back as the front surface of your body becomes convex and the back surface of your body becomes concave (an arch into the back space). As you sit down allow your pelvis to sink and retreat toward the seat with your head coming forward and your front surface becoming convex while your back surface becomes concave.
2. Practice Breath Cycle Out/in Even: After establishing your clapping rhythm try exhaling on a count of three claps and inhaling on a count of three claps.
- Imagine you have just taken a freshly baked cookie out of the oven. Exhale to blow on the cookie to cool it down and then inhale to savor the aroma of the cookie.
 - To heighten your experience, as you exhale allow your head to come forward toward your hand while allowing your core to retreat so that your front surface becomes concave. As you inhale, condense your core while lifting your head and rising into the zone of high as you savor the aroma.
3. Practice Breath Cycle In/out Uneven Long/short: After establishing your clapping rhythm try inhaling on a count of four claps and exhaling on a count of two claps.
- Imagine you are outside on top of a mountain with a beautiful vista in front of you. As you inhale breathe in the beauty of the endless vista. Then a mosquito bites the front of your neck and take a short exhale as you slap it.
 - To heighten your experience, as you take a long inhale, allow your arms to sweep to the side and then overhead as your core rises and spreads open to breathe in the endless vista. As your arms arrive overhead the mosquito bites your neck. Take a short exhale as your hand slaps your neck in order to squash the mosquito. As you do this, allow your core to retreat and sink while you become concave in your chest around your hand on your neck.
4. Practice Breath Cycle Out/in Uneven Short/long: After establishing your clapping rhythm try exhaling on a count of two claps and inhaling on a count of four claps.
- Imagine you are brushing out your long hair. As you begin the brushstroke you encounter a knot. Take a short exhale as you yank the brush through the knot and then take a long inhale as you finish the brush stroke unimpeded by tangles.
 - To heighten your experience, allow your hand to condense around the brush as you encounter the knot and take a short exhale with some force and suddenness. As you inhale and continue the unimpeded brush stroke, allow your arm to arc to the back as you tilt your head back pulling the brush through the long hair. Allow your chest to rise and advance as the front surface of your body becomes convex.

5. Practice Breath Cycle In/out Uneven Short/long: After establishing your clapping rhythm try inhaling on a count of two claps and exhaling on a count of four claps.
 - Imagine you are typing on your computer and you are startled by a loud noise. Gasp, taking a short inhale as you turn your head toward the sound. Realize that it is an acorn falling on your metal roof and take a long exhale as you relax and calm down.
 - To heighten your experience, as you inhale allow your body to condense and turn toward the sound as your hand comes quickly up to your mouth. As you exhale allow your hand to release to your side as you retreat and sink back into your chair allowing your body to expand into the seat.
6. Practice Breath Cycle Out/in Uneven Long/short: After establishing your clapping rhythm try exhaling on a count of four claps and inhaling on a count of two claps.
 - Imagine you are sitting on the floor leading story time to a group of children sitting in a circle at the library. As you exhale quiet the children with a “shhh” sound. Then take a short inhale to prepare to begin to tell the story.
 - To heighten your experience, as you exhale put your finger across your lips, allowing your core to tilt forward toward the children and make an arc with your core (leading with the head) that traverses the circle. As you inhale rise to an upright seat to begin your story.

12.5 Choreographer, Performer, and Audience Analysis

In this section, we analyze three moments from *Babyface*¹ that correspond to three of the previously identified breath patterns from the audience, performer, and choreographer perspective to show how emotion, visual aesthetic, and style can work together to create a meaningful moment. Breath is a channel of volitional and unconscious control. Humans need to breathe to survive, but we also use it to support bodily action as well as an expressive channel itself.

Moreover, as in engineering design, our choreography had to include choices about where performers should put their energy and resources based on what the choreographers, designers, and performers think an audience will observe and understand. Likewise, the machine design was a resource-limited construction (and we can't detect breath perfectly in the resultant system). There are both artistic and analytical design choices to be made, particularly here where the choreography is entwined with a machine the performer wears that is designed to detect changes in the bodily volume associated with breathing.

¹ The timestamps given in the figure captions align with the video of the performance available at: <https://vimeo.com/kateladenheim/babyface>.

We can, then, think of the components of style, aesthetics, and emotion as a collaboration between creators and audiences. In *Babyface*, the choreographer works inside a futuristic visual aesthetic to create a stylized femme character who embodies a cyborg existence. Performers enact different styles of breath patterns to move the machine they are wearing in a way that corresponds with the desired choreography. This concert of human and machine movement brings together an onstage character that may resonate emotionally for members of the particular Western, digital culture in which the work was made.

Figure 12.3 shows the distinction between the complexity of the performer's overall action and the machine's action, which is created from one aspect of the performer's bodily movement: the changing volume in their torso. Similarly, the simplicity of the breath strategy and resulting machine movement (shown in the center and rightmost motifs, respectively) relative to the complexity of the overall bodily action (leftmost motif) is seen in the analysis in Figs. 12.4 and 12.5.

Figure 12.3 highlights a moment where the application of breath to motion works to contrast an aspirational state (a sense of upright tension, especially in the spine) with a motion that sends the upper body rolling backwards in an attempt to communicate a sense of failure, exhaustion, or giving up. The complexity of this bodily production is shown in the leftmost motif: reading from bottom to top, a becoming phrase of expansion and convexity, modified by concavity in the backspace and sustained time effort, is followed by a still shape form of a wall, modified by sudden time effort, to end the impactive phrase. In the center and rightmost motifs we see the simple breath strategy that produced a bipart phrase in the machine as well. However, the machine movement does not have the same

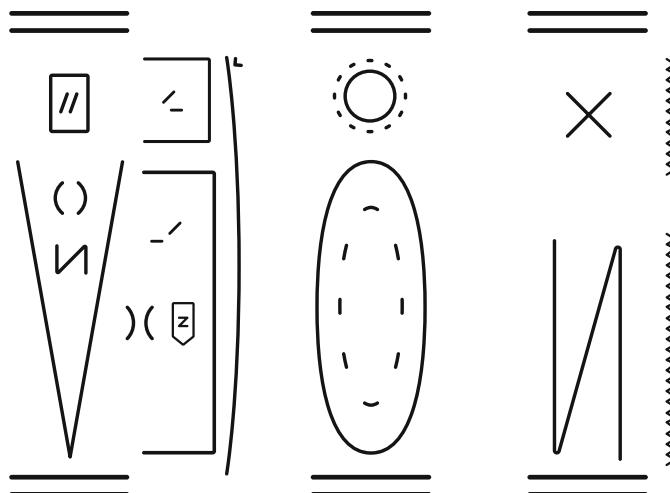


Fig. 12.3 Motif for Breath Pattern 1 (3:34–4:00); audience observation (left), performer strategy (center), and machine action (right). An effort to communicate a contrast between aspiration and failure

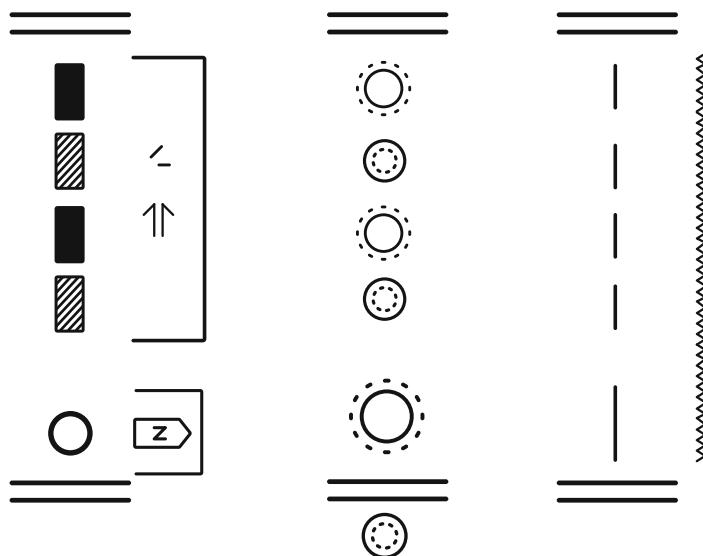


Fig. 12.4 Motif for Breath Pattern 3 (2:05–2:07); audience observation (left), performer strategy (center). and machine action (right). Choreographer vision statement: An exaggerated mime of a courtesy giggle, with punctuated rhythm in the hands, shoulders and hips

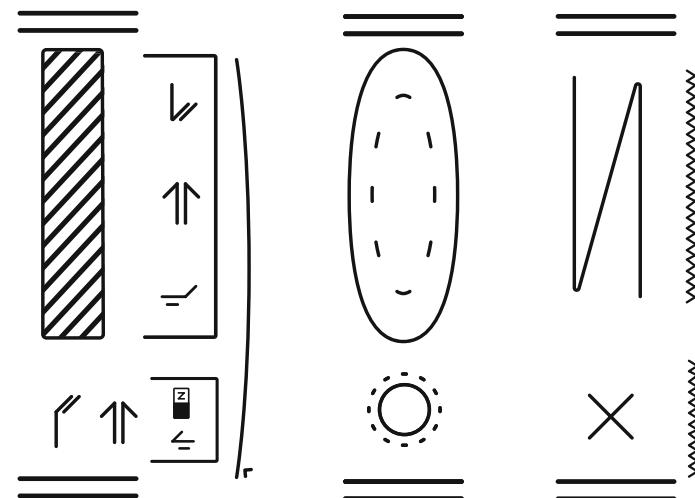


Fig. 12.5 Motif for Breath Pattern 5 (0:00–0:05); audience observation (left), performer strategy (center). and machine action (right). Choreographer vision statement: This is a moment that is meant to communicate focus, preparation and control and to visually set up the mechanic of breath to robotic motion

integrated coherence as the performer's breath, as indicated by the blank space and vibratory phrasing bow around the basic body actions of expand and contract.

Figure 12.4 focuses on an application of breath that is couched within a rhythmic, campy section accompanied by driving hyper-pop music. It's an exaggerated mime of a cutesy giggle, and the breath pattern reflects this punctuated rhythm in the hands, shoulders and hips. The leftmost motif, reading top to bottom, shows a simple side hold followed by four up and down actions of the shoulder girdle modified by quick time. Yet, underlying these are several short breaths, including a preparatory inhale, which happens below the start of the motif, that create the vibratory "liveness" of the mechanical wings through the breath sensor. This machine action is barely read as "action", not rising the level of expansion and contraction as in the other motifs.

Finally, Fig. 12.5 shows a moment at the very outset of the piece, that is about focus and control. A long, intentional inhale that inflates the torso is the sole focus of the performer, instructed by the sound score, and reflected in machine choreography. It's meant to call attention to the beginning of the piece, and equate an inhalation with expansion, grandeur, and beauty. Here, as shown in the leftmost motif, an action of sinking shape quality releases to the upward dimension inside an impulsive phrase. As shown in the center motif, the performer uses a short exhale followed by a long inhale that expand the wings, setting up for the audience in this opening moment the mapping between her movement and the robotic wings as the narrator says "take a deep breath."

12.6 Relating Breath to Robots in Public Spaces

This work offers practical guides to how breath can relate to our experience of machines as well. Associating inhale with expansion and presentation and exhale with contraction and turning inward is a stylistic choice, which sits inside an aesthetic created by plastic textures, purple and green hues, and metallic lines. This creates a design that allows for the performer's expression to read more clearly for audience members who have likely seen such futuristic aesthetics crafted from happy, helpful robots, which contrasts the emotive state—dejected failure—created by the choreography and performance of *Babyface*.

Robots in public spaces will need similarly designed interactions, connecting our embodied experience to the affordances of the machine. These devices need stylistic choices about movement design and aesthetic ends that create the physical features of the machine, which will inform how the devices create or interact with internal and emotive states of human counterparts. For example, a machine that is meant to guide tourists through a museum space will need the design of a motion trajectory through space for a "guiding" action for a particular degree of freedom from somewhere on the device, which will either succeed in clarifying the way for human counterparts or create confusion and distress. Either of these outcomes may be evident in the rate of breath of the guided visitor, which can be measured using wearable devices, e.g., heart rate monitors, in a mapping that is specific to this context (in another context, such as an exercise class, an increased breathing

rate may be a desirable outcome). Thus, artwork like *Babyface*, utilizing somatic experience as a tool for research, becomes an important site for investigating our own bodies, designing the movement of machines, and bridging their interaction.

In applying artistic methods to robotics design, the authors do not wish to suggest that simplistically making robots breathe is always an appropriate approach. Breath is a rich activity with myriad interpretations, reflecting vast emotional states and variations in conscious and semi-conscious control. We do not recommend machine design in which breath patterns are mapped to inflexible motions and presumed emotive outcomes. We do recommend using breath as part of a choreographic, designerly and interactive toolbox, which offers a more nuanced look at the action and how it facilitates human interaction with machines.

As a choreographic and somatic tool, breath facilitates the execution of movement. Filling and emptying the body of air has a direct impact on the body's shape and abilities, therefore impacting actual and perceived sensations of fluidity, tension, effort and ease. These changes influence how a performer communicates with their body, which can include influencing emotional states in themselves and for an audience. A skilled choreographer notices and takes advantage of these performative qualities, directing them towards various functional and expressive ends. We see this as a place where dance as an academic field, and particularly choreographic expertise, can support motion design in artificial systems.

Such choreographic tactics are already widely used in character animation, video games, and virtual environments. These systems use breathing animations and gentle, oscillating inward/outward motions to offer a body (abstract or humanoid) illusions of liveness. For robots, such animations might similarly indicate an “on” state, or that the machine is functioning properly. Breakdowns in breathing regularity or labored affectations, as modeled in *Babyface*, might be used as a convention to indicate that the machine is malfunctioning or needs repair.

12.7 Conclusion

This chapter has outlined the role of visual aesthetic, style, and emotion in the interactive performance piece *Babyface*. We have discussed how the visual aesthetic of the choreographer provides a vision statement and medium in which a work is produced; moreover, we have seen how different styles of movement are enacted by performers to create onstage phenomena that resonate, often emotionally, with audiences.

For robotic design similar choices will affect the perceived style, aesthetics, and emotions of these devices. In this chapter, we use *Babyface* to investigate this particular design-space around breath in order to suggest features of breath that robots in public spaces may need to model in order to function effectively. To this end, we've highlighted a range of applications in technology where breath may offer fruitful explorations.

Acknowledgments The authors would like to thank the organizers and attendees of the HRI in Public Spaces workshop at the ACM/IEEE International Conference on Human-Robot Interaction (HRI) 2022 where an earlier version of this book chapter appeared as a workshop paper. Regarding the creation of *Babyface*, the authors would like to thank Reika McNish and Wali Rizvi for technical and creative support in preparing for this engagement. We thank Ken Cooley and his team at ShapeMaster for fabrication. We also thank Myles Avery for music and soundscape design as well as production support in New Zealand. This showing of *Babyface* received generous support from the U.S. Embassy in New Zealand. Curator Sam Trubridge and his full team at The Performance Arcade in Wellington New Zealand offered production support throughout conception and installation. We also want to thank Footnote New Zealand Dance (General Manager Richard Aindow and Artist Liaison Anita Hunziker) and company members Oliver Carruthers, Sebastian Geilings, Rosie Tapsell and Cheyenne Teka who performed during the 5 day engagement. We also thank Colin Edson and Eric Minnick, who assisted with construction and provided production support during this event. This work was funded by a grant from the U.S. Embassy in New Zealand, performance fees from Dance NOW and the Conference for Research on Choreographic Interfaces (CRCI), and support from the Robotics, Automation, and Dance (RAD) Lab as well as the Mechanical Science and Engineering and Dance Departments at the University of Illinois at Urbana-Champaign.

References

1. Bradley, H., Esformes, J.D.: Breathing pattern disorders and functional movement. *Int. J. Sports Phys. Ther.* **9**(1), 28 (2014)
2. Corness, G., Schiphorst, T.: Performing with a system's intention: embodied cues in performer-system interaction. In: Proceedings of the 9th ACM Conference on Creativity & Cognition, pp. 156–164 (2013)
3. Cuan, C., Pakrasi, I., Berl, E., LaViers, A.: Curtain and time to compile: a demonstration of an experimental testbed for human-robot interaction. In: 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 255–261. IEEE (2018)
4. Cuan, C., Pakrasi, I., LaViers, A.: Time to compile. In: Proceedings of the 5th International Conference on Movement and Computing, pp. 1–4 (2018)
5. Elswit, K.: Reflections on bodies in lockdown: the coronasphere. *Multimodal. Soc.* **1**(1), 69–74 (2021)
6. Gemeinboeck, P.: The aesthetics of encounter: a relational-performative design approach to human-robot interaction. *Front. Robot. AI* **7**, 577900 (2021)
7. Hackney, P.: Making Connections: Total Body Integration Through Bartenieff Fundamentals. Routledge (2003)
8. Herath, D., Jochum, E., St-Onge, D.: The art of human-robot interaction: Creative perspectives from design and the arts. *Front. Robot. AI* **9**, 910253 (2022)
9. Hutchinson Guest, A.: Motif at a Glance. Language of Dance Centre, London (2000)
10. Laban, R., Lawrence, F.: Effort: A System Analysis, Time Motion Study. MacDonald & Evans, London (1947)
11. Ladenheim, K., LaViers, A.: Babyface. In: Proceedings of the 7th International Conference on Movement and Computing, pp. 1–2 (2020)
12. Ladenheim, K., LaViers, A.: Babyface: Performance and installation art exploring the feminine ideal in gendered machines. *Front. Robot. AI* **8**, 576664 (2021)
13. Ladenheim, K., McNish, R., Rizvi, W., LaViers, A.: Live dance performance investigating the feminine cyborg metaphor with a motion-activated wearable robot. In: 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 243–251. IEEE (2020)
14. LaViers, A.: Counts of mechanical, external configurations compared to computational, internal configurations in natural and artificial systems. *PLoS One* **14**(5), e0215671 (2019)

15. LaViers, A.: Ideal mechanization: Exploring the machine metaphor through theory and performance. *Arts* **8**(2), 67 (2019)
16. LaViers, A.: Dancing with robots. *Am. Sci.* **108**(4), 236–240 (2020)
17. LaViers, A.: First encounters with robots through embodied observation, imagined narrative and choreography. In: *Materializing Digital Futures: Touch, Movement, Sound and Vision*, p. 169 (2022)
18. LaViers, A., Maguire, C.: *Making Meaning with Machines: Somatic Strategies, Choreographic Technologies, and Notational Abstractions through a Laban/Bartenieff Lens*. MIT Press (2023)
19. Matheus, K., Vázquez, M., Scassellati, B.: A social robot for anxiety reduction via deep breathing. In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 89–94. IEEE (2022)
20. Nettl-Fiol, R., Vanier, L.: *Dance and the Alexander Technique: Exploring the Missing Link*. University of Illinois Press (2011)
21. Saegusa, R., Ito, H., Duong, D.M.: Human-care rounds robot with contactless breathing measurement. In: *International Conference on Robotics and Automation (ICRA)*, pp. 6172–6177. IEEE (2019)
22. Stedge, C.: *Breath*. Laban/Bartenieff Institute of Movement Studies [Unpublished thesis for CMA Certification] (2017)
23. Studd, K., Cox, L.: *Everybody is a Body*. Dog Ear Publishing (2019)
24. Terzioglu, Y., Mutlu, B., Şahin, E.: Designing social cues for collaborative robots: The role of gaze and breathing in human-robot collaboration. In: *15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 343–357. IEEE (2020)
25. Tran, Q.V., Su, S.F., Nguyen, V.T.: Pyramidal Lucas–Kanade-based noncontact breath motion detection. *IEEE Trans. Syst. Man Cybern. Syst.* **50**(7), 2659–2670 (2018)

Chapter 13

Humanist-in-the-Loop: Machine Learning and the Analysis of Style in the Visual Arts



Kathryn Brown

Abstract This chapter examines theories about artistic style and explores the extent to which they can be used support the computational modeling of painted surfaces. It is argued that style comprises more than the visible features of paintings and that this poses specific challenges to the algorithmic analysis of artworks. Beyond acts of categorization, can computer vision challenge fundamental ideas about style and its role in art historical analysis? What kinds of questions do computer scientists need to ask in order to create meaningful data sets and to undertake stylistic analyses of the works contained in them? How should statistical anomalies be dealt with and do they have significance beyond their mere outlier status? In proposing answers to these questions, the chapter defends a contextual approach to the analysis of style. It is argued that close collaboration between computer scientists and art historians is required in order to embed the results of quantitative analyses in relevant socio-cultural frameworks and to move beyond the instrumentalization of existing formal or morphological models.

13.1 Introduction

One of the advantages of using computational methodologies to analyze style in the visual arts is the possibility of making feature comparisons across a large sample of artifacts. The works in question may be drawn from different periods of an individual artist's career or may be associated with creative production undertaken more broadly in particular geographies, schools, or time periods. Computer vision analysis of large data sets is capable of offering new insight into attribution, genre, and technique among other matters because it facilitates work on a scale that cannot typically be undertaken by an individual researcher [15, 18, 20]. Yet making this kind of quantitative study meaningful in humanities contexts requires

K. Brown (✉)
Loughborough University, Leicestershire, UK
e-mail: K.J.Brown@lboro.ac.uk

the careful integration of algorithmic and art historical approaches to the gathering and interpretation of data.

The aim of this chapter is to relate the computational modeling of painted surfaces to important theories about artistic style. Specifically, I will be concerned with what “style” means, how it signifies, and which elements beyond the visible features of the art object fall within the parameters of the term. The discussion is motivated by the following questions: What kinds of questions do computer scientists need to ask in order to create meaningful data sets and to undertake stylistic analyses of the works contained in them? In its use of statistical methods to perform acts of categorization and analysis, does computer vision exacerbate or challenge problems about the use of stylistic analysis in art history? How should statistical anomalies be dealt with and do they have significance beyond their mere outlier status? My aim is not only to examine the potential of computational methodologies in the visual arts, but also to suggest how such methodologies prompt a reflection on the concept of “style” itself. For reasons of space, I will focus on the use of computer vision to analyze style in paintings.

13.2 Debating Style

There is a large body of secondary literature about style in both art history and analytic aesthetics, and it is beyond the scope of this chapter to give a detailed account of the wide-ranging definitional questions and debates to which the term has given rise. For the purpose of this chapter, it is worth noting, however, that by the late 1970s, the American art historian George Kubler suggested that “style” had become “a word to avoid” on the grounds that it was simply “gray with fatigue” [9, p. 163]. Like many scholars, Kubler was concerned with problems that had historically arisen from uses of the term and its reduction to “near-formlessness” in debates about category problems in the visual arts [9, p. 168]. He suggested—somewhat colorfully—that attempts to determine the dominant qualities of art production in specific time periods had led researchers to “treat historic styles as though they were persons in a generational novel” [9, p. 164].

Kubler’s essay highlighted the existence of conceptual hierarchies that derived from attaching markers of style to particular time periods, places, and methods of working. Writing in a similar vein, art historian Svetlana Alpers elaborated her own concerns about the “radically historical bias” encountered in this field [1, p. 137]. For Alpers, the study of style runs the risk of “extracting, by naming and singling out, the accomplishment of specific modes that seem by virtue of this nomination to have preeminence” [1, p. 158]. The identification of “style” becomes, she argues, a self-fulfilling prophecy: “style is what you make it” [1, p. 158]. Alpers’s preference is to dispense with attempts to identify and emphasize specific features that are used to denote individual or period style and to adopt, instead, a “modal” way of thinking about the various links that connect makers, their works, and the world [1, pp. 158 and 162]. While this may seem like style by another name, Alpers was

concerned with the wedge that studies of style seem to drive between the artist and the wider environment in which an individual's varied acts of making and imagining are embedded. According to Alpers, a modal approach "has the virtue of not distinguishing form and content, of not excluding function, of not choosing in advance between the parts played by the individual maker, his community, certain established modes of perceiving the world, or the viewer" [1, p. 158].

Alpers's attempt to understand style as encompassing more than an abstract (or even arbitrary) set of visible properties belonging to one or more art objects is an idea that has also been explored in analytic aesthetics. Carolyn Wilde, for example, argues that the style of a painting is "a method of looking that is publicly constituted through the materials of art" [21, p. 124]. Wilde acknowledges and agrees with the points raised by Alpers and, accordingly, draws into her discussion the impact of the materials that artists use as well as the imaginative and social transactions that impact on the making and reception of their works. In a similar approach that integrates different elements of making, seeing, and understanding ways in which pictures relate to the world, Paul Crowther [6] breaks down style into a series of functions. He suggests that style is an activity expressed in (at least) four ways: (1) the choice of medium and facility in using it; (2) decisions as to subject matter (including which parts of a subject on which to focus); (3) composition; and (4) relationship to the style of others. Crowther's definition captures the idea that making images is a "learned competence" [6, p. 7] which relates in different ways to historical and prevailing creative practices, handling of materials, and culturally determined ways of seeing.

The approaches to style discussed above privilege openness to a range of ideas pertaining to creativity that extend beyond the visible characteristics of the artwork. The analysis of relationships between artist and public, individual expression, and social context motivated Michael Baxandall's conception of the "period eye" in his discussion of fifteenth-century Italian painting [2, pp. 29–57]. In his use of this formulation, Baxandall sought to capture the patterns, categories, inferences, and analogies that a viewer brings to an image for the purposes of deriving meaning from it [2, p. 34]. These range from sensitivity to an artist's use of specific colours to the appreciation of narrative backgrounds for particular scenes, recognition of the signifying power of gestures, and ways in which forms relate to each other. The ability of audiences to recognize and understand the meaning attaching to such elements of a painting (all of which may be said to pertain to the work's style) comprise, for Baxandall, ways of seeing that mark out a distinct historical "period". In similar fashion to Alpers's discussion, form and content are understood to be intimately connected. According to Baxandall, a painter in fifteenth-century Italy utilized a specific range of tropes in the production of works for audiences who shared a set of experiences and expectations. The result, he argues, is that the visual encounter with an image constituted "a marriage between the painting and the beholder's previous visualizing activity on the same matter" [2, p. 34]. Here too, therefore, "style" is a term that requires a broad conception of the physical, social, and imaginative elements that comprise the work of art.

The acts of shared understanding that underpin Baxandall's notion of the "period eye" echo a distinction made by the philosopher Richard Wollheim in his exploration of artistic style. Wollheim offered a detailed discussion of the topic in *Painting as an Art* [22] and further refined key ideas in his essay "Style in Painting" [23]. Wollheim's account has spurred much discussion in aesthetics [19], and I will only mention certain aspects of it here. Wollheim drew a distinction between two broad concepts of style: "general style" and "individual style". The former concerns the characteristics of a School of painting or the qualities of works produced during a particular historic period and the latter relates to the distinctive qualities of an individual artist's production, including the ways in which those qualities may change over time.

In Wollheim's discussion, individual style has intrinsic relevance to the artist, is the product of that individual's psychology, and it is something that is acquired by the artist over time. This part of the discussion generates a further distinction. Wollheim suggests that individual style sub-divides into two capacities: the first involves "segmenting or conceptualizing the elements of a painting in a certain preferred way" to identify the "schemata" of an artist's style. The second consists in the artist "evolving rules or principles for operating with these schemata" by, for example, giving shape to them on a painting's support [22, p. 42]. For Wollheim, these two capacities are exercised by the artist (they are practical as opposed to theoretical) and they are dependent on context (style might result in outputs that look different from each other). Wollheim is keen to point out that his view is not a "formalist" account of style. "Formalism", he argues "ought to think of schemata as narrowly configurational items which a careful scrutiny of the support will reveal inscribed on it" [23, p. 43]. In contrast to this, Wollheim offers a deliberately capacious account of the elements that comprise the schemata of an artist's style, including the material qualities, represented elements, and figurative aspects of a painting that fulfill the artist's intentions [23, pp. 43–44].

Wollheim's focus on "segmentation" and the examination of rules-based schemata uncannily anticipates the kind of language that computer scientists use when determining how to approach and divide the surface elements of a canvas for the purpose of algorithmic analysis. On the basis of the discussion in this section, however, it should be clear that there are tensions in determining what elements of a painting are included in such segmentation, how visual sense might be made of the open-ended or generative elements of an artist's schemata, and the extent to which wider contextual elements of art production and reception factor into the notion of style. In the following section, I will explore how these issues impact on computer vision analyses of paintings and will debate some of the problems and possibilities to which they give rise. While Lev Manovich claims that "computers are always more precise in their descriptions of characteristics of analog cultural artifacts" [11, p. 1146], my argument is that there is more to style than meets the eye.

13.3 Humanist-in-the-Loop

In a study of what machine learning can contribute to the study of artistic style, Elgammal et al. [7] debated how algorithms can be developed to classify style, what internal representation might be used to achieve this, and how the results might relate to methodologies in art history. For the purpose of their analysis, the authors adopted a famous approach to the study of artworks by Heinrich Wölfflin published in 1915 *The Principles of Art History: The Problem of the Development of Style in Early Modern Art*. This entailed developing deep learning models to test classifications that Wölfflin proposed in his writings about certain changes that occurred in European painting styles from the fifteenth to the seventeenth century.¹

Wölfflin's work is a foundational text and has spurred interesting debates in both art history and aesthetics, but the classificatory apparatus posited in *The Principles of Art History* is not one to which art historians would typically turn today [16], [12, p. 3]. Wollheim points out that in purporting to identify general style in "overtly configurational terms" [23, p. 45], Wölfflin gave the impression that matters of an artist's individual style could be derived from larger schemata and on similar terms. While there is scientific interest in operationalizing the categories found in works such as those of Wölfflin, the results of such an analysis are limited in their capacity to open new avenues in the interpretation of paintings. Nuria Rodríguez-Ortega has noted the risk of producing "a sort of neoformalism or neovisualism in the field of art history", an approach that occludes a range of meanings "that exceed the merely visual" [15, pp. 349–350]. What kinds of cross-disciplinary conversation need to take place in order to harness the innovations of computer vision without simply replicating such formalist or morphological models?

I would make the case that a contextual approach to the analysis of style is likely to achieve insightful results. As discussed in the opening section, this necessitates close attention to a range of factors (e.g., social, physical, and material) that impact on the development of an artist's individual repertoire of gestural expression and on the social and interpretive transactions that make the resulting bundle of features identifiable bearers of meaning. It will be clear from the above, that I am arguing for the need to keep humanists (and, indeed, the concept of the viewer) in the loop while developing models for the computer vision analysis of paintings. Remaining attentive to the issues that Alpers, Wilde, and Wollheim raise about the handling of materials and the social transactions that allow both artists and audiences to derive meaning from images, the bare isolation and extraction of features from the surface of a work does not, in itself, give an accurate picture of what might be described as either general or individual styles of painting. Similarly, operationalizing historical

¹ The five pairs of concepts that Wölfflin explored were: linear vs painterly; plane vs recession; closed vs open form; multiplicity vs unity; absolute vs relative clarity. For further discussion see Nanay [13, pp. 149–150]. The other key historical reference point in discussions of style is Alois Riegl's 1894 treatise *Problems of Style: Foundations for a History of Ornament* [14].

models of style runs the risk of omitting more recent and important art historical and aesthetic debates about factors that shape the very concept of style.

When considering which features of a painted surface to extract and compare, it is necessary to consider how materials and contexts influence the creation and interpretation of datasets. Wollheim describes the physical components of a painting as “elements that depend upon the artist’s materials not only for their realization [...] but also for their identification” [22, p. 43]. This encompasses, for example, types of pigment (oil paint, oil stick, acrylic paint, pastel blends, watercolors, etc.) and their methods of application. It also includes less obvious factors that impact on the shape of marks made using such materials, including the features of an image’s support (e.g., the priming and weave of a canvas or the blend of fibers that comprise a piece of paper). While it might be possible to identify and isolate evidence of an artist’s repeated gestural work across a number of paintings, other material and environmental factors impact on the way in which those gestures are executed or appear on different occasions.

Computer vision is highly effective in isolating the visual details of paintings, but it is also necessary to contend with statistical outliers within larger patterns that emerge from the application of quantitative methodologies. A case where an artist has accidentally overloaded a paintbrush or where environmental elements have impacted the surface of a work during its creation (e.g., the presence of grains of sand in Claude Monet’s *The Beach at Trouville*, 1870, National Gallery London) create anomalies that might be discounted from stylistic trends determined by computer vision analyses. From an art historical point of view, however, such statistical outliers can be as interesting as the dominant results derived from the application of quantitative methodologies. As an illustration of Claude Shannon’s theory of information, the unexpected or surprising result derived from the computer analysis of a group of paintings may, in fact, yield a larger amount of information than the more statistically probable outcome [17, p. 188]. By identifying outliers and debating their relationship to dominant patterns, the combined work of computer scientists and art historians may be able to offer more specific insight into the uses of materials in specific contexts and to determine how an individual’s approach to mark making might have been affected by matters that are not typically taken into account when determining the hallmarks of a particular style.

This leads to issues concerning the ways in which an artist’s style is understood within the trajectory of an individual’s lifetime. Important work has been undertaken in pattern recognition and statistical analyses of surface features of paintings to provide insight into different periods of Vincent van Gogh’s creative career and ways in which his techniques of paint handling differ from those of his contemporaries [8, 10]. In addition to providing insight into matters of attribution, such studies connect productively to Wollheim’s idea that before forming a style, an artist produces “pre-stylistic” work and that, over the course of a career, further differences may ensue including the production of “post-stylistic” or “extra-stylistic” work [22, p. 42]. I am generally sympathetic to Wollheim’s account of style, but this aspect of his discussion (and the kinds of computer vision analyses to which it might give rise) can be difficult to accommodate in the case of some artists.

The creative practice of Henri Matisse is a good example. During the early decades of his career, Matisse experimented with Realism (*Woman Reading*, 1894, Musée national d'art moderne, Paris), Impressionism (works painted at Belle-Île, France in 1896–1897) and Divisionism (*Luxury, Calm, and Pleasure*, 1904, Musée d'Orsay, Paris). He used these styles deliberately and his pursuit of them perplexed audiences when the resulting works were first shown. Critics, viewers, and gallerists alike sought (and failed) to identify a distinctive “Matissien” style that the artist employed systematically and consistently [24, pp. 21–25] [4, pp. 46–60]. It is difficult to characterize such works as pre- or extra-stylistic, and, indeed, Matisse went on to create works in numerous different styles over the remainder of his career. In order to make sense not just of the changes that emerge over the course of an individual's creative life (and hence any dataset derived from them), it is necessary therefore to be sensitive to the broader personal, creative, and social factors that impact on aesthetic decisions made by the artist.

Inevitably, discussions about style in art history lead to debates about connoisseurship and the extent to which computer vision might productively replace the subjective biases associated with trained looking by humans. “Connoisseurship” is a term that has fallen out of use in art history. In part, this is because of the exclusivity, power, and privilege that it has historically connoted as well as its contribution to the creation of hierarchies of value endorsed by art markets [5, p. 159]. If, as Mansfield et al. suggest, connoisseurship “is now rarely viewed as an art historical end in and of itself but rather as subsidiary to historical interpretation and analysis” [12, p. 3], it is legitimate to ask why computer scientists would seek to teach connoisseurship to “a new kind of observer—the computer” [3, p. 1].² On the holistic account of style that I have explored in this chapter, it is unlikely that a computer could, in fact, replace an expert (whether or not identified as a “connoisseur”) as there is more to an understanding of style than “learning from examples” or developing “strategies of attribution” [3, p. 1]. As in the case of the algorithmic instrumentalization of existing models to analyze form, the replacement of human vision by machine learning in such cases leads to a loss of expert input and failure to account for factors that lie beyond the purely visual.³

The question that arises—and this is the driving question of this chapter—is how the broader features that fall within the concept of style can be successfully incorporated into computer vision analyses of artworks? While algorithmic analyses isolate specific, visible features of paintings, these analyses need to be augmented by art historical expertise to embed the quantitative results in meaningful notions of what constitutes “style”. This requires conceptual agility on the part of both computer scientists and art historians. If, as Alpers [1] suggests, it is preferable to avoid reducing matters of style to features that have been pre-determined as important before the analysis of a painting is made, researchers need to adopt a

² For an opposing view of the role of the connoisseur see [5, p. 161].

³ See also Rodríguez-Ortega on the value of intellectual collaboration in the field and the absence of art historians from much discussion about computer vision in the arts [15, pp. 351–352].

flexible approach to how a painting might be segmented, how and to what end machines may be trained to analyze data sets, and how the results are interpreted. While this might entail revisiting and revising the computer vision analysis of paintings during the course of a study, it also means that researchers must remain open to the merits of informational surprise and, hence, to the value of not pre-determining forms and functions. Computational analysis can produce a vast amount of important data about paintings—much of which is inaccessible to the human eye. However, it is only by keeping the humanist in the loop that the full potential of such analyses can be realized.

References

1. Alpers, S.: Style is what you make it: The visual arts once again. In: Lang, B. (ed.) *The Concept of Style*, revised and expanded edition edn., pp. 137–162. Cornell University Press, Ithaca (1987)
2. Baxandall, M.: *Painting and Experience in Fifteenth-Century Italy: A Primer in the Social History of Pictorial Style*. Oxford University Press, Oxford (1988)
3. Bell, P., Offert, F.: Reflections on connoisseurship and computer vision. *J. Art Historiography* (24), 1–10 (2021)
4. Brown, K.: *Henri Matisse*. Reaktion Books (2021)
5. Carrier, D.: In praise of connoisseurship. *J. Aesthetics Art Crit.* **61**(2), 159–169 (2003)
6. Crowther, P.: *Defining Art, Creating the Canon: Artistic Value in an Era of Doubt*. Clarendon Press, Oxford (2007)
7. Elgammal, A., Liu, B., Kim, D., Elhoseiny, M., Mazzone, M.: The shape of art history in the eyes of the machine. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, pp. 2183–2191 (2018)
8. Kotovenko, D., Wright, M., Heimbrecht, A., Ommer, B.: Rethinking style transfer: From pixels to parameterized brushstrokes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12196–12205 (2021)
9. Kubler, G.: Toward a reductive theory of visual style. In: Lang, B. (ed.) *The Concept of Style*, revised and expanded edition edn., pp. 163–173. Cornell University Press, Ithaca (1987)
10. Li, J., Yao, L., Hendriks, E., Wang, J.Z.: Rhythmic brushstrokes distinguish Van Gogh from his contemporaries: Findings via automated brushstroke extraction. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(6), 1159–1176 (2012)
11. Manovich, L.: Computer vision, human senses, and language of art. *AI Soc.* **36**, 1145–1152 (2021)
12. Mansfield, E., Zhang, Z., Li, J., Russell, J., George, S., Young, C.A., Wang, J.Z.: Techniques of the art historical observer. *Nineteenth-Century Art Worldwide* **21**(1), 125–136 (2022)
13. Nanay, B.: Two-dimensional versus three-dimensional pictorial organization. *J. Aesthetics Art Crit.* **73**(2), 149–157 (2015)
14. Riegl, A.: [1893] *Problems of style: Foundations for a History of Ornament*, edited by David Castriota; translated by Evelyn Kain. Princeton University Press, Princeton, NJ (2018)
15. Rodríguez-Ortega, N.: Image processing and computer vision in the field of art history. In: Brown, K. (ed.) *The Routledge Companion to Digital Humanities and Art History*, pp. 338–357. Routledge, New York (2020)
16. Schwartz, F.J.: Cathedrals and shoes: Concepts of style in Wölfflin and Adorno. *New German Critique*, (76), 3–48 (1999)
17. Stone, J.V.: *Vision and Brain: How We Perceive the World*. MIT Press, Cambridge, MA (2012)

18. Stork, D.G.: Computer vision and computer graphics analysis of paintings and drawings: An introduction to the literature. In: Computer Analysis of Images and Patterns: 13th International Conference, CAIP 2009, Münster, Germany, September 2–4, 2009. Proceedings 13, pp. 9–24. Springer (2009)
19. Van Gerwen, R. (ed.): Richard Wollheim on the Art of Painting: Art as Representation and Expression. Cambridge University Press, Cambridge (2001)
20. Wang, J.Z., Kandemir, B., Li, J.: Computerized analysis of paintings. In: Brown, K. (ed.) The Routledge Companion to Digital Humanities and Art History, pp. 299–312. Routledge, New York (2020)
21. Wilde, C.: Style and value in the art of painting. In: Van Gerwen, R. (ed.) Richard Wolheim and the Art of Painting, pp. 121–134. Cambridge University Press, Cambridge (2001)
22. Wollheim, R.: Painting as an Art. Thames and Hudson, London (1987)
23. Wollheim, R.: Style in painting. In: Van Eck, C., McAllister, J., Van de Vaal, R. (eds.) The Question of Style in Philosophy and the Arts, pp. 37–49. Cambridge University Press, Cambridge (1995)
24. Wright, A.: Matisse and the Subject of Modernism. Princeton University Press, Princeton (2004)

Part V

Aesthetics

In this part, we explore how modern computational technologies can enhance the aesthetics and overall quality of images.

The first chapter, “The Inter-relationship between Photographic Aesthetics and Technical Quality,” looks at how the technical quality of an image influences its aesthetics and vice-versa. It reviews existing literature, compares the use of deep learning features for both quality assessment tasks, and provides a fresh perspective.

The second chapter, “Image Restoration for Beautification,” focuses on image restoration—a process to improve the aesthetics of degraded images. It covers key topics and shares the latest research results in this field.

Concluding this part is the chapter “Image Affect Modeling: An Industrial Perspective.” The chapter explores how image affect modeling has been used in the industry and presents insightful discussions on practical applications and future opportunities.

Chapter 14

The Inter-Relationship Between Photographic Aesthetics and Technical Quality



Franz Götz-Hahn, Lai-Kuan Wong, and Vlad Hosu

Abstract Quality assessment has become increasingly important, finding applications in image and video acquisition, synthesis, search, and more. Image quality assessment (IQA) and image aesthetics assessment (IAA) are two types of quality assessment techniques that are related but have distinct objectives. IQA focuses on technical quality, such as low-level defects, while IAA considers abstract and higher-level concepts that encompass the subjective aesthetic experience. Recent IQA techniques have broadened their scope, emphasizing the interactions between the perception of defects and aesthetics. However, few studies have investigated the relationship between photographic aesthetics and technical quality. The interconnection between the two facets of quality is apparent subjectively, from an empirical and theoretical perspective, as well as objectively, as their connection influences the cross-task prediction performance of IQA and IAA models. We review the literature identifying shared factors considered in models for IQA and IAA, compare the use of deep learning features in each task, and discuss other connections at a theoretical level.

F. Götz-Hahn
Universität Kassel, Kassel, Germany
e-mail: franz.goetz-hahn@uni-kassel.de

L.-K. Wong
Multimedia University, Cyberjaya, Selangor, Malaysia
e-mail: lkwong@mmu.edu.my

V. Hosu (✉)
Konstanz University, Konstanz, Germany
e-mail: vlad.hosu@uni-konstanz.de

14.1 Introduction

Technical quality and aesthetics are both kinds of visual quality. The two have been studied separately until recently. Technical quality is essential for image and video acquisition, processing, storage, and display systems. Consequently, quality has been seen as the lack of impairments and is related to visual fidelity. The quality of experience has extended this definition to include the viewer's context and personal influences such as mood and emotions. On the other hand, aesthetics has long been in the realm of artists. Initially studied in a non-technical context, it has been recently approached from a technical standpoint. When it comes to evaluating visual stimuli, quality focuses on the degree of annoyance caused by distortions, while aesthetics takes into account the emotional impact it has on the viewer as a whole.

Technical and aesthetic quality share underlying psychological and perceptual mechanisms. We explore their relationship in the context of the human visual system. At an abstract level, quality and aesthetics may seem separate, with quality emphasizing the form of the visual message, while aesthetics focuses on the emotional impact of the message as perceived by the viewer. However, the two are intertwined. This is because, in the human visual system, both bottom-up and top-down effects are at play simultaneously. Thus, quality, a characteristic of the form of the message, influences aesthetics, which is an effect of the interpretation of that message. The same goes in reverse; the final interpretation affects the perception of the form and, thus, its perceived technical quality. Consequently, a study of aesthetics can benefit from understanding the connections with technical quality and vice versa.

14.1.1 Overview

Computational assessment methods for aesthetics and technical quality typically predict subjective image ratings. This is done by training machine learning models on annotated datasets. Modern techniques have been greatly improved by the advent of deep learning, enabling more accurate predictors. Nonetheless, this comes at the cost of reducing the interpretability of the features used. Despite these advances, the modeling of aesthetic quality remains more challenging than that of technical quality. Reasons for this difference include the highly subjective nature of aesthetics, variations in modeling techniques, and the reliability and diversity of the datasets used. We can enhance our understanding of aesthetic modeling by examining the connections between quality and aesthetics. Thus, we explore the commonalities in more detail.

The interdependence between quality and aesthetics is apparent in the scientific literature, which often addresses aspects that are associated with both. When

evaluating photographs, certain attributes can either detract from or enhance the intended message of the work. For example, image noise, blurs, poor contrast, or overexposure can degrade the image's quality and, thus, the fidelity of the message. In contrast, intentional restyling or enhancement, as well as the use of degradations that evoke nostalgia or suggest authenticity, e.g., movie grain, can improve the image's aesthetic appeal by altering the emotional impact of the message. Thus, the same factors can be perceived contextually as either degradations or enhancements.

Despite the apparent connections between image quality and aesthetics, few attempts have been made to quantify this relationship. The most direct measure of agreement is the correlation between aesthetics and quality scores on the same set of images, rated preferably by the same users. However, the reliability of datasets containing both aesthetics and quality scores [86, 88] is limited. Drawing any definitive conclusions is challenging due to the inconsistent correlation values found for each dataset, which range from a low of 0.14 SRCC (0.17 PLCC) on PaQ-2-PiQ [88] to a high of 0.98 SRCC (0.99 PLCC) on PARA [86]. It is likely that each correlation is not representative of the level of agreement we would achieve in a well-designed experiment, either due to psychological biases, insufficient controls (environment, attention), or imprecise task definitions. For instance, in the PaQ-2-PiQ dataset, the quality scores for the 9081 images shared with the AVA dataset are annotated in a separate experiment from the aesthetics scores, which come from AVA. The different levels of reliability, and distinct participants, lead to an extremely low agreement. The quality and aesthetics scores in the PARA dataset, on the other hand, are provided for each image by the same participants and in immediate succession—each viewer rates the quality of an image and then, within a few seconds, does the same for aesthetics. This may have introduced an anchoring bias, which led to similar quality and aesthetics scores being assigned to the same image [27]—the median SRCC per user is 0.9, even though the study is using a 9-point discrete scale for aesthetics. Together with other contextual factors, such as the small presentation size of the images, the two tasks are conflated and become relatively indistinguishable from each other. It is expected for there to be a near-perfect agreement between the aesthetics and technical quality scores in this scenario.

As many datasets are reliably rated for either technical or aesthetic quality and not both, we can only compare their agreement by creating machine learning models and testing them cross-dataset. We take this approach later to get a better understanding of the connection between the two tasks.

Next, we review the current literature on feature-based modeling and do an in-depth analysis of the use of deep features in quality assessment. Early works have focused on creating handcrafted features related to subjective factors for either aesthetics or quality, enabling us to examine the types of features used and their significance. Deep features, which have proven to be valuable in transfer learning, are also studied to determine their potential for perceptual assessment. In our discussion of the relationship between quality and aesthetics, we also explore other theoretical aspects in addition to the aforementioned topics.

14.1.2 Background

Research specifically dedicated to investigating the relationship between the aesthetic and the technical quality of images is scarce. However, several articles have found connections between the two. Namely, Cerosaletti and Loui [5] researched the effects of a selection of image characteristics on perceived technical and artistic properties. The authors studied the presence and size of a person in the image, overall image sharpness, image composition, and other concepts, relating them to the aesthetic appeal ratings. They performed principal component analysis on the distributions of aesthetic appeal ratings and identified clusters of images that had similar levels of quality and types of quality degradations present. Essentially, they found that the first principal component, which highly correlated to mean artistic ratings, also aligned with the presence, location, and severity of technical image defects. Their analysis suggested that the images in the study formed a continuum both along the image aesthetics and image quality axes, simultaneously. Furthermore, they concluded that it is advisable to initially filter for sufficient technical quality when trying to identify aesthetically pleasing images.

Redi and Heynderickx [69] considered the “integrity” of an image—a quasi-synonym for technical image quality—and compared it to its aesthetic appeal. They selected a set of high technical quality images as a high-integrity reference dataset and created visibly degraded JPEG-compressed versions that were grouped as a corresponding low-integrity dataset. Participants of a subjective study then assessed the aesthetic appeal of the resulting images of both classes. Additionally, they were instructed not to take into account an image’s integrity when judging the aesthetics. In their analysis, the authors found no statistically significant difference in aesthetic appeal when comparing the two datasets against each other. However, within the low image integrity dataset, aesthetics and technical image quality scores were positively correlated, implying some relationship. The integrity of low technical quality images was also negatively correlated with the difference in aesthetic appeal between the datasets of high and low technical quality images. This suggests that increasing the level of distortion also decreases the perceived aesthetics of an image, supporting the earlier results from Cerosaletti and Loui [5]. This finding is also supported by previous work [78, 79] that examined the effects of image quality manipulations on the aesthetic appeal of photographs. The results also showed a clear preference for high-quality scenes over manipulated ones.

Finally, within the group of high technical quality images, Redi and Heynderickx [69] found an inverse relationship between aesthetic appeal and technical quality scores. The more aesthetically pleasing an image was perceived to be, the lower the image scored in terms of technical quality. Overall, these findings suggest a complex relationship between image quality assessment (IQA) and image aesthetics assessment (IAA).

14.2 Interconnections in Representations

The ability to effectively represent data and capture relevant semantics is essential in the application of machine learning techniques. In the fields of IQA and IAA, expert domain knowledge has traditionally been used to design handcrafted feature representations, which involves a time-consuming engineering process. Despite their differing goals—IQA is concerned with measuring the degree and presence of distortions, and IAA evaluates the aesthetic experience and sense of beauty derived from an image—these tasks share common image factors, such as color, contrast, and blurs.

Recently, researchers in both fields have operationalized the tasks by designing systems that automatically learn feature representations, which have the potential to improve the accuracy and generalization ability of models in both IQA and IAA. Notably, previous works in the two fields have adopted similar convolutional neural networks (CNNs) as the backbone for prediction, indicating common deep feature representations shared between the tasks. The following section will discuss the within-domain state-of-the-art for image quality and aesthetics, providing an upper limit for the cross-test performance. We use cross-task and cross-domain interchangeably and refer to changes in the images' source domain and the corresponding rating types: aesthetics and technical quality. Then, we investigate the empirical relationship between quality and aesthetics using cross-domain testing and analyze the value of particular features for the individual tasks.

14.2.1 Traditional Feature Representations

Image factors used in traditional IQA and IAA prediction can be divided into three categories: distortion-related, low-level visual features, and high-level semantic features. Distortion-related features carry no semantic meaning by themselves. They mainly describe the presence, location, and level of particular distortions primarily caused by an element in the image capture, storage, or transmission pipeline. Low-level features are simple characteristics that can be extracted from an image without taking shape information or spatial relationships into account, and they contain limited semantic information. From a psychological standpoint, these low-level features might be similar to those processed in the first 100 ms by the human visual system [38]. They primarily provide a global description of the image, such as the overall distribution of brightness and colors. High-level features, on the other hand, are concerned with finding shapes and objects, as well as spatial relationships between objects that describe the semantics of image contents. Features in this category relate to conceptual representations of the scene and objects within it. Table 14.1 contains an inexhaustive list of image factors from all three categories with references to their use in IQA and IAA.

Table 14.1 Relevant previous work that used traditional features to model image factors that help in predicting types of quality assessment

	Image factors	IQA	IAA
Distortion/Artifacts	Noise	[1, 17, 50, 67] [10, 39, 92, 93]	
	Blockiness	[2, 6, 46, 47, 90] [7, 8, 39, 49]	
	Ringing	[7, 8, 48, 95]	
	Blurring	[9, 15, 59, 60, 80] [7, 21, 49, 92, 93]	[11, 81, 83]
		[8, 10, 39, 64]	
Low-level visual features	Contrast/lighting	[36, 66, 82] [44, 53, 64]	[11, 30, 81, 83]
	Color	[53, 64]	[11, 30, 56, 81, 83]
	Texture	[53]	[11, 83]
High-level semantic features	Shape		[11, 81]
	Composition		[11, 23, 28, 56, 83]
	Subject features		[56, 83]
	Subject-background contrast		[83]

From Table 14.1, we can observe that factors used in technical quality prediction fall mainly in the distortion and low-level visual features categories. Earlier IQA methods focused on detecting common spatial distortion using image processing techniques (e.g., wavelet, discrete cosine transform, histogram analysis, and filtering), but recently the emphasis has been placed on making the quality estimation more in line with the quality perceived by the human visual system [71]. In general, traditional methods can be categorized into approaches that focus on a singular factor and those that consider multiple factors at the same time. Single factor methods act as a detector to recognize and measure specific types of distortions, e.g., spatial noise [1, 17, 50, 67, 68], blockiness [2, 6, 46, 47, 90], ringing [48, 95] and blurring [9, 15, 21, 59, 60, 80]. As an image often contains several artifacts, some researchers employ multiple factors for quality assessment, in which a suitable mechanism is used to combine the results of different artifact measurements to estimate the overall perceptual quality. Zhu and Milanfar [92, 93] estimate blur and noise, Liu et al. [49] measure blockiness and blur, and Chetouani and Beghdadi [7] predict quality based on blur and ringing. In addition to blockiness and blur, Li [39] as well as Chetouani and Beghdadi [8] also estimate noise and ringing, respectively, for quality prediction.

Apart from distortion-related factors, some IQA methods consider low-level visual features, e.g., contrast, color, and texture. Peli et al. [66], and Winkler and Vanderghenst [82] measure contrast using filtering, while Lai and Kuo [36] use wavelets. Lin et al. [44] compute contrast in terms of luminance and Lu et al. [53]

extend it to include color-contrast and texture-contrast. An alternative approach is taken by Ouni et al. [64], who train an artificial neural network with a set of semantic features consisting of measures for sharpness, clarity, brightness, and colorfulness. In the video quality domain, a similar approach is pursued by Men et al. [61].

The subjective nature of IAA presents a challenge, which may explain why there are only a few models designed with handcrafted features, unlike IQA. The representation of aesthetic features can be broadly classified into low-level and high-level visual features. The only distortion factor utilized in aesthetics modeling is blurring [11, 81, 83]. As lighting is a key to aesthetically pleasing photographs, all traditional IAA models represent lighting factors in the form of brightness, exposure, or contrast [11, 30, 56, 81, 83, 87]. Colors are represented in various forms such as color properties (e.g., hue and saturation) [11, 83, 87], colorfulness [11, 87], color distribution [30] and color harmony [56]. Texture is commonly represented using wavelet filter banks [11, 83, 87].

High-level semantic factors such as shape, image composition, subject-focused features, and subject-background contrast are among the different aesthetic factors used in IAA. Shape properties such as curvature, convexity, and diagonal lines are often linked to the aesthetic intent of the photographers. Several methods successfully use shape [11, 81, 87] as features for IAA. Image composition—the interplay of image elements in a photograph—has its aesthetic root in the visual arts. Early handcrafted IAA methods model only a subset of simple composition rules, such as simplicity [56, 83, 87], rule-of-thirds [56, 83], visual balance [28] and fill the frame [83]. Recently, more holistic approaches were introduced to model and analyze composition. Li et al. [40] exploit spatial design categories and the Notan Dark-Light principle to build a composition feedback mechanism to provide suggestions on composition and tonal enhancement respectively. Based on the triangle concept used by professional photographers to compose their photographs, He et al. [23] design an algorithm to automatically detect embedded triangles for analyzing the composition of portraits.

The subject of a photograph can greatly impact its overall quality, which is why photographers often employ a range of techniques to emphasize the subject and make it dominant in the frame. Such techniques may include creating contrast between the subject and the background, using appropriate framing, and applying a shallow depth-of-field. Several researchers [30, 56, 83] extract low-level features representing the subject. In their work, Wong and Low [83] leverage subject features to calculate the subject-background contrast, a novel feature that serves as a measure of subject dominance. Notably, modeling these high-level aesthetic factors is a highly complex task as it requires semantic scene understanding. Moreover, a particular composition rule may not apply to all categories of photographs; for instance, the rule-of-thirds is applicable for a wide shot scene, but may not be suitable for a close-up portrait. Undoubtedly, designing a holistic set of aesthetic feature representations is a challenging task.

The review of empirical features in Table 14.1 reveals that both technical and aesthetic quality share common representations, specifically blurring and low-level features, namely contrast/lighting, color, and texture. Interestingly, although IQA

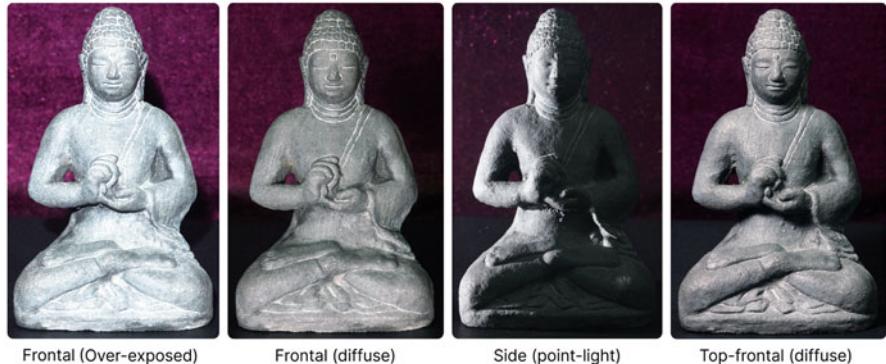


Fig. 14.1 Influence of light source placement on contrast. We show various types of lighting that affect the aesthetics and technical quality of a composition. After ranking the images subjectively (mean of three observers), we ordered them from the lowest aesthetic quality (left) to the highest (right). Image © Vlad Hosu [24]

and IAA share common factors, the measure of these factors may have a different scale of sensitivity and complexity. Lighting is a good example to illustrate this point. Lighting creates a 2D pattern of contrasts that helps the brain recognize 3D objects. IQA focuses on contrast as a degradation caused by the limitations of the acquisition device in extreme lighting conditions. The resulting photographs show a loss of detail, regions of overexposure with washed-out or dull colors, or harsh distracting shadows. In IAA, on the other hand, contrast is the orchestrated interplay of tones and colors. When appropriately designed, it can direct the viewer's attention and enhance the aesthetic experience. In this sense, the measures of contrast in IQA and IAA have different levels of sensitivity and complexity, with IQA focusing on technical factors and IAA emphasizing abstract interpretation as a tool to manipulate perception.

Figure 14.1 illustrates the effect of different lighting placements and light sources on image contrast. The images are arranged by their aesthetic appeal, from the lowest on the left to the highest on the right (based on a limited subjective study), with the top-frontal (diffuse) having the highest IAA score. In terms of technical quality, we can observe that the frontal (over-exposed) and side (point-light) have poor contrast, which means these two images would have lower IQA scores than the other two. Interestingly, the frontal (diffuse) image looks flat and uninteresting, although the contrast is not unappealing. Conversely, while some details on the right side of the “point-light” image are lost due to harsh shadows, aesthetically, it portrays a striking contrast and utilizes a “shadow-play” technique that professional photographers frequently employ to capture captivating and visually appealing photographs.

Blurs resulting from unintended camera shake or long exposures are considered a crucial factor in determining technical quality ratings, often leading to low scores. Nonetheless, the link between blur and aesthetics is not as straightforward. There



Fig. 14.2 Intentional blurs that enhance aesthetic appeal. The first row shows the use of motion blur to add a sense of movement, while the second row uses bokeh to emphasize and enhance the subject. Images are made available on Unsplash.com from their respective authors

are several forms of blur that photographers intentionally create to achieve specific effects that can enhance a work's visual appeal. For example, the top two images in Fig. 14.2 show how motion blur can be used to create the appearance of movement and speed. The biker is clearly visible and sharp in the photograph on the right, but the background is strongly blurred, creating the illusion of movement. The bokeh effect is an artistic technique applied to emphasize the subject of a photo by blurring the background, as illustrated in the bottom two images in Fig. 14.2. Conventional DSLR cameras can easily capture this effect, but mobile cameras face challenges due to limitations in their image optics.

These examples demonstrate the fascinating interplay between IQA and IAA. In most cases, good technical quality is a fundamental prerequisite for aesthetics—an image with poor technical quality is often unappealing. However, their relationship has its intricacies, for instance, some images with low technical quality may possess high aesthetic value.

14.2.2 Deep Feature Representations

In state-of-the-art approaches to both IAA and IQA, the use of large-scale deep neural networks has become the norm. Deep networks that have been pre-trained on different tasks, such as image classification, are a starting point for fine-tuning or a source of rich features. Table 14.2 lists state-of-the-art IQA and IAA methods,

with their respective choice of deep neural networks used for extraction of deep feature representations. Unsurprisingly, similar convolutional neural networks have been used for both technical and aesthetic quality predictions. Earlier works adopted AlexNet [34] or self-crafted shallow CNNs inspired by AlexNet. Later works employed deeper networks such as VGG-16 [73], Inception [76], or ResNet [22]. At this point, it is not an overstatement to say that deep neural networks' internal representations of useful features have effectively replaced traditional handcrafted features for both IQA and IAA tasks.

The best-performing models in the field of technical image quality are based on pre-trained convolutional neural networks [3, 26, 74], which involve automatically learned hierarchical features. Similarly, in aesthetics prediction, classical feature-based approaches covering characteristics related to color, composition, and simplicity are no longer compared, since models using deep features explicitly [25] or implicitly [52, 72, 77] outperform them significantly. The complexity of manually designing high-level feature predictors for perceptual attributes like beauty is partly responsible for this shift.

To provide some insights into the effectiveness of deep features, we examine the visualization of the learned deep feature representations. The feature visualization at different depths of an Inception-v1 network [70] in Fig. 14.3—obtained by maximizing the activation of a specific kernel in a particular layer—shows how the network builds up its “understanding” of images over many layers. Alternatively, the visualizations can be understood to be a representation of what creates the strongest activation of an individual kernel. It is evident that the features in the early layers activate on specific alignments of edges and textures. Activation of features in the middle part of a network relates to entities with simple semantic meanings, such as patterns and object parts. Further up the network, individual features might activate on very specific objects of increasing complexity, such as a building or a dog’s face. Finally, the latter layers show a more abstract representation that seems to combine features from previous layers.

From this visualization, we can observe that a deep neural network can extract a more holistic feature set for improving prediction performance. Analogous to

Table 14.2 Relevant previous work that used features extracted from pre-trained deep neural networks to model both types of quality assessment

Deep architecture	IQA	IAA
Shallow CNN	[29, 32, 57]	[54, 55, 58]
AlexNet	[3, 65]	[33, 94]
VGG-16	[4, 16, 43, 51]	[35, 77]
ResNet18	[88]	[72, 86, 91, 94]
ResNet50	[74, 84]	[35, 86]
InceptionResnetV2	[26, 45]	[25]
InceptionV2/V3		[42, 77, 94]
MobileNetV2		[77, 86]
DenseNet121		[42]
SwinTransformer		[86]

quality assessment, earlier layers can be effective for capturing distortions/artifacts. In contrast, middle layers can capture more perceptual features that carry some semantic meaning, e.g., lighting and color harmony—this corresponds to the low-level visual features from Table 14.1. Finally, the upper layers capture high-level features related to image composition. Moreover, we can see that individual features are tailored toward individual classes of objects, which, in theory, can render them more powerful than traditional features.

A deeper dive into IQA research shows that pioneering deep IQA methods [29, 32] divide an image into patches and train self-crafted shallow CNNs to extract features from all patches, which are then combined to regress an image's quality score. Following that, Bosse et al. [4] and Liu et al. [51] use the deeper VGG-inspired CNN for feature extraction. As computer capacity expanded, increasingly complicated neural networks with multiple sub-networks or multiple tasks have been used for IQA prediction. Lin and Wang [43] introduce the hallucinated-IQA, which jointly optimizes three sub-networks: the quality-aware generative network, the hallucination-guided quality regression network, and the IQA-discriminator. Ma et al. [57] and Yan et al. [85] present multi-task deep neural networks that predict distortion type and natural scene statistics in addition to IQA prediction. DeepBIQ [3] and PaQ-2-PiQ [88] built deep region-based architectures that perform IQA prediction on multiple sub-regions of the original image. More recently, several works proposed deep neural networks that capture multiple levels of feature representations to improve IQA prediction. HFD-BIQA network [84] extracts the local visual structure via an orientation selectivity mechanism, and combines it with high-level semantic features extracted with ResNet. BLINDER [16] extracts deep features from each layer of a pre-trained VGG-net. DeepFL [45] took inspiration from Hosu et al. [25] in the IAA domain to go a step further by utilizing the deeper architecture InceptionResNet-v2 to extract and aggregate features from multiple levels, at their original size (512×384 pixels) to retain detailed information.

The development of IAA models shares a similar trend as those introduced for IQA. Earlier methods employ shallow networks, and the latter become more

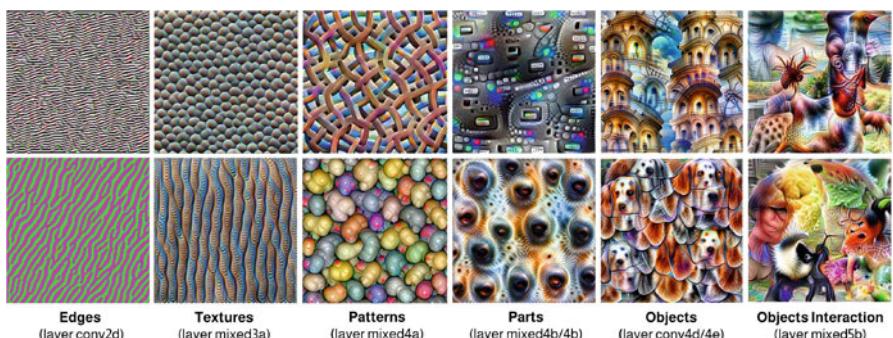


Fig. 14.3 Visualization of deep visual features from different layers in the Inception-v1 network. Image © OpenAI, taken from the OpenAI Microscope

complex, capturing a more comprehensive set of features related to distortions and varying abstraction levels of object categories. Compared to early deep learning IQA methods that use single-column shallow CNNs, RAPID [54] and DMA-Net [55] present slightly more complex architectures that adopt double- and multi-column shallow CNNs to extract global-local and multi-patch features respectively. Kong et al. [33] propose an AlexNet-based deep architecture with several augmentations that unify aesthetics attributes and photo content for image aesthetics ratings. Mai et al. [58] present a composition-preserving deep ConvNet method that combines the predictions from multiple sub-networks with adaptive spatial pooling sizes by leveraging a scene-based aggregation layer. Talebi and Milanfar [77] introduce the NIMA architecture that utilizes a pre-trained CNN model, with augmented FC and softmax layers for predicting aesthetics. Hosu et al. [25] extract features from all convolutional blocks of a pre-trained InceptionResNet-v2 network to exploit multi-level representation at full image resolution and train a custom shallow CNN on these new features—called multi-level spatially-pooled (MLSP) features. Motivated by the subjectivity of aesthetic preference among individuals, several researchers in the field of IAA have started to explore personalized aesthetics assessment techniques [42, 86, 94]. More recently, Farhat et al. [14] explore the idea of personalized aesthetics-aware image retrieval to assist photographers in capturing high aesthetics photos, by-passing the need to assess aesthetic quality from an image.

Even though IQA and IAA have primarily been investigated individually, the interconnection between these two quality aspects is apparent from the theoretical and empirical comparison of feature representation and predictive modeling. This interconnection would potentially influence the cross-domain test performance of image quality and aesthetics models. The fact that identical source networks have been employed for various tasks related to perceptual attributes implies that the features used for these tasks hold significant information for both. This has been touched upon briefly by Talebi and Milanfar [77], who introduced cross-task testing, where the models trained for one task were tested on the other. The current model optimization approach only focuses on individual tasks, without considering how the relationship between IQA and IAA could improve the accuracy of both tasks’ predictions.

14.3 Prediction

As previously stated, the goal of both image aesthetics and visual quality prediction is the design of computational models whose predictions strongly correlate with the subjective evaluation of human observers. Classically, this is done within a particular domain, which is usually constrained to a single dataset at a time. The dataset is split into training and test sets, where the model’s performance on the test set represents an unbiased estimate of the generalization performance of the model in that particular domain. Since datasets differ in terms of image sources and

annotators, or even the phrasing of the question presented to the assessors, a model's performance can vary greatly when tested on data from a domain other than the one it was trained on. Consequently, the image quality community, in particular, has recently put a greater emphasis on cross-testing, meaning that the trained model is evaluated on data from a different dataset than what it was trained on. The argument for this evaluation is easily understood, as the utility of a trained model lies in its generalization to data from different sources. In the following, we will describe the within-domain state-of-the-art for both the image quality and image aesthetics domains, which provides a baseline or best-case performance. Then, we investigate the empirical relationship of the domains using cross-domain testing and analyze the value of particular features for the individual tasks.

In our evaluations, we investigate the utility of deep features. Specifically, we use MLSP features extracted using the approach proposed by Hosu et al. [25]. Figure 14.4 depicts the extraction process, where global average pooling is applied to the activations of all kernels in the Stem, each Inception-A, Inception-B, and Inception-C module, as well as the Reduction-A, Reduction-B, and Reduction-C modules, resulting in a 16,928-dimensional vector of features. At the time of writing, this feature representation is one of the best sources of information for IAA and IQA tasks at comparable dimensionality. Although the best performance on MLSP features has been achieved using neural networks, we perform our evaluations using the XGBoost library, as it allows us to derive feature importance for a prediction task using built-in mechanisms.

14.3.1 Within- and Cross-Domain Testing

For this investigation, we will focus on the KonIQ-10k [26] and AVA [63] datasets as representatives for the image quality and image aesthetics domains, respectively. First, KonIQ-10k is the largest IQA dataset sampled for diversity and reliably annotated by crowdsourcing. The top contender to KonIQ-10k would be SPAQ [13],

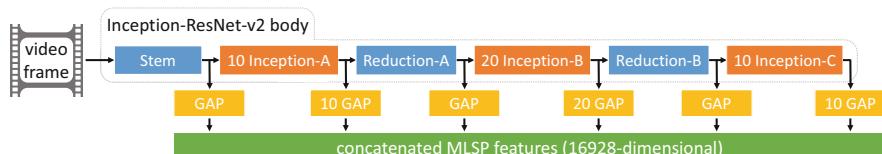


Fig. 14.4 Extraction of multi-level spatially-pooled (MLSP) features from an image, using an InceptionResNet-v2 model pre-trained on ImageNet. The features encode quality-related information, with the initial layers capturing finer details of the image such as its sharpness or noise, whereas the subsequent layers act as object detectors or encode information pertaining to visual appearance. Global average pooling (GAP) is applied to the activations resulting from the Stem, each Inception-module, as well as the Reduction-modules, and finally concatenated to form MLSP features

which, although $\approx 10\%$ larger and annotated in lab conditions, contains only smartphone photos. We select KonIQ-10k as it enables the best generalization performance when testing on other datasets [75] potentially even beyond technical quality assessment, i.e., aesthetics. Second, AVA is the largest and most popular benchmark dataset for IAA. Recently proposed aesthetics datasets such as PARA [86] are not yet established enough to consider as a baseline, and appear to be not as reliably annotated.

When using the official KonIQ-10k test set in a within-domain testing scenario, our XGBoost models achieve a baseline of 0.908 SRCC, which is roughly comparable to the state-of-the-art works that show performances between 0.916 and 0.938 SRCC [26, 31, 41, 74, 75]. For all further evaluations, we use 80% of the data as a training set, leaving 10% for validation and testing, respectively, for KonIQ-10k. To make the performance comparisons and feature evaluations fair, we limit the data used for training on AVA to the absolute size for KonIQ-10k. This means roughly 8000 images are used for training, while 1000 images are used for validation and testing, respectively, yielding an SRCC of 0.585 on AVA. For reference, using all available data to train on, our model achieves 0.706 SRCC on the official AVA test set, which is close to state-of-the-art [25, 31, 62, 89]. We believe this setup enables informed conclusions about deep features' predictive utility.

These baseline performances indicate a stark difference between the two datasets. Despite the amount of data used being similar, the model performances differ by over 0.3 SRCC. There are multiple potential reasons for this, as already previously alluded to. It is possible that the fundamental concept of aesthetics is more difficult to be derived from the features used in this scenario. Alternatively, the interpretation of an image's aesthetic appeal might be more subjective. Different annotation protocols for the datasets have led to unreliable ratings for AVA.

14.3.2 Feature Importance

Nonetheless, using XGBoost's built-in feature importance values, we can investigate the utility of individual kernels in the MLSP vector for the prediction of the two tasks. "Feature importance" is calculated for a single decision tree by the amount that a node, i.e., each split in the tree, improves the performance measure, weighted by the number of observations the node was considered in. Finally, these values are averaged across all trees, to obtain the overall feature importance. As the name suggests, it essentially encodes the contribution an individual feature provides to a particular prediction task. Figure 14.5 shows a histogram of feature importance for the two datasets. The feature importance of the IQA task is roughly one order of magnitude higher than for image aesthetics, which could be an indication that the features do not capture characteristics of aesthetic appeal well. Additionally, we show the average joint rank on the right, which shows that a particular subset of features has an extraordinarily low rank. Here, the joint rank is calculated as the average index of corresponding kernels in the MLSP vector when sorted by

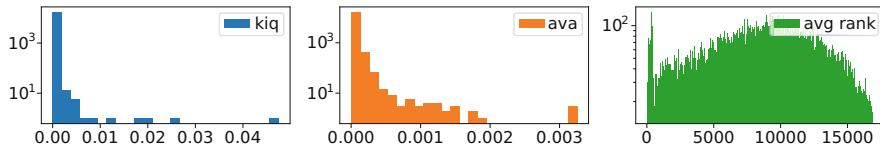


Fig. 14.5 Histogram of feature importance for KonIQ, AVA, as well as the average rank of all features when trained on each task. For both of the individual tasks, only a few individual features are considered important by the XGBoost models. This is evident both in the two individual histograms but also in the fact that a small set of features has a particularly low rank, meaning their respective importance is especially high

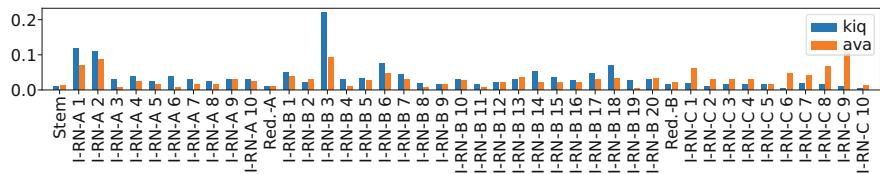


Fig. 14.6 Histogram of top 1% features according to feature importance grouped by the Inception module they reside in for IAA and IQA tasks, as well as jointly

importance values in descending order. These approximately 500 features represent generally useful features for both tasks.

Another interesting aspect is the location within the Inception-ResNet-v2 parent network, where the most useful features reside for each of these two tasks. Figure 14.6 shows a histogram of features according to their source modules for the top 500 MLSP features for the individual tasks. Particular blocks seem to be more relevant for the prediction of both perceptual attributes. The “I-RN-A 1” and “I-RN-A 2” blocks, which are both located at the bottom of the network and represent very simple filters, are two examples of that. On the other hand, some modules are more relevant for the prediction of one of the two tasks. Most notably, modules towards the head of the network are assigned higher importance for the IAA task, while the network modules in the bottom and midsection score higher in terms of feature importance for IQA. This observation supports the hypothesis that image quality is primarily governed by features captured in the early parts of a network that describe low levels of abstraction. On the other hand, factors that play a role in judging image aesthetics are found frequently towards the head of a network, which is much more abstract in nature, as shown in Fig. 14.3.

As previously stated, cross-testing is an increasingly used evaluation method for the generalization performance of models. However, this commonly refers to testing on datasets within the same task. Since we are interested in the connections between two different fields, image aesthetics and image quality, we evaluate the cross-test performance between the two tasks. Consequently, we evaluate the models trained on AVA by testing on KonIQ-10k test sets and compute the SRCC between the predicted (aesthetics) scores and the ground truth quality scores, and vice versa. The

aesthetics predictors evaluated on IQA achieved a correlation of 0.473 SRCC, while the quality predictors evaluated on IAA achieved an SRCC of 0.261. This amounts to approximately a 50% performance drop caused by changing the prediction task. In previous works, the cross-test of an aesthetics assessor trained on AVA was reported to be 0.6 SRCC for KonIQ-10k. The IAA method itself achieved 0.76 SRCC on AVA. The ratio between the performance observed for the two cross-testing directions was 0.79 ($= 0.6/0.76$). Notably, the XGBoost approach produces a similar ratio of 0.81 ($= 0.473/0.585$). This suggests that the relative performance of the machine-learning approach is more important than its absolute performance values. Thus, even though XGB's performance is somewhat inferior to state-of-the-art IAA models, it is still representative in relative terms.

14.3.3 Feature Selection

Since we are interested in finding features that are informative for both tasks, we investigate feature selection methods aiming to retain performance in both domains. As a baseline, we are comparing the performance when selecting the same number of features at random. We sampled 100 random feature combinations of $n \in \{1, 4, 16, 66, 264, 1058\}$ features and subsequently trained an XGBoost model on a unique 90/10 train/validation split with the test set held out. The resulting performances evaluated on the official test sets for KonIQ-10k and AVA are visualized as estimated contour plots for the joint probability densities in Fig. 14.7. Additionally, we plotted the top feature combinations in the same color, obtained by taking the same number of features with the highest importance values of the

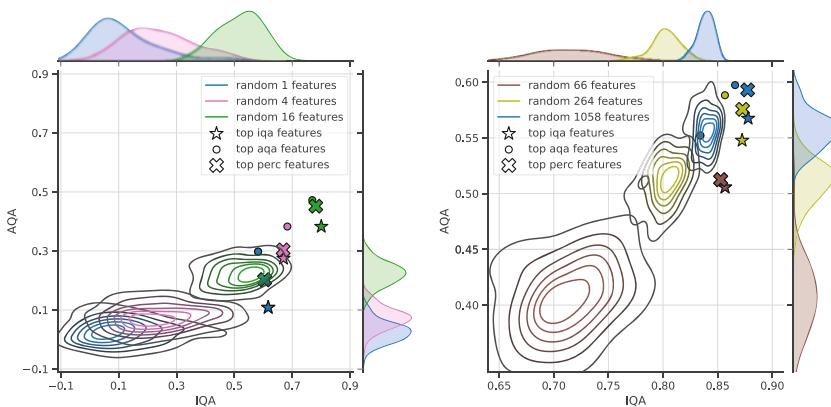


Fig. 14.7 Performance of XGBoost models on KonIQ-10k and AVA when trained on different numbers of features taken either randomly (density plots) or as the top features according to individual joint feature importance on both tasks (' \times ' markers). For the density plots, we took 100 random feature combinations

individual tasks, as well as the joint importance. The joint importance, denoted by the crosses, was computed as the product of the individual task importance values.

Any of the top feature combinations outperform random selections by a significant amount. In fact, using any of the three most informative feature combinations exceeds the average performance of a randomly selected set of four times as many features. Moreover, models trained on features selected by considering one of the two tasks generally perform better in that task. When considering the tasks simultaneously, the performance difference is bridged to some extent. Nonetheless, objectively evaluating performance improvement is not trivial. A naïve way could be to take the Euclidean norm as $\sqrt{\rho_{\text{IQA}}^2 + \rho_{\text{IAA}}^2}$, where ρ indicates the SRCC on the respective dataset, and comparing the difference between the mean performance of a randomly sampled feature set to that of the most important perceptual features. However, this difference does not represent a fair measure, as an improvement in this norm has to be evaluated within the context of the baseline performance.

The significance of a selection of features improving a baseline of $\rho_{\text{IQA}} = \rho_{\text{IAA}} = 0.2$ by 0.05 SRCC is lower than the impact of a selection strategy that improves a baseline of $\rho_{\text{IQA}} = \rho_{\text{IAA}} = 0.8$ by the same amount. Nevertheless, the improvement according to the Euclidean norm described above would seem larger, relatively speaking, for the former case. An alternative way to approach the problem of measuring improvements fairly would be to consider the reduction in the Euclidean norm from some theoretical maximum performance, e.g., $\rho_{\text{IQA}} = \rho_{\text{IAA}} = 1$. Although this performance may be unreasonable or unachievable, using it as an anchor point to measure the improvement would be more interpretable.

14.4 Discussion

The results of our previous investigation demonstrate that the proposed approach can be an exciting avenue to find and understand features that are useful for both aesthetics and quality prediction. However, the assessment of image quality and aesthetics involves multiple factors, which are not always easy to identify and quantify. In this section, we discuss the factors involved in quality and aesthetics judgments. We explore the differences between technical and subjective factors and how they relate to the procedures devised for taking better photos. We also delve into the complexity of these factors and how to measure it in principle. By analyzing the factors that contribute to quality and aesthetics, we can better understand the challenges that arise in image assessment and how to tackle them.

14.4.1 On Factors Involved in Quality and Aesthetics Judgments

Quality and aesthetics assessment are both affected by multiple factors. Quality factors stem largely from types of artificial distortions introduced during the acquisition, processing, and display of images. Thus, factors for quality are technical and concrete, as well as often related to side-effects of the application of known algorithms, such as upscaling or demosaicing blurs and compression artifacts. Aesthetic factors are more subjective, not having as clearly identifiable technical causes. They relate to rules of thumb or principles that are used by photographers. Thus, technicality is related to the procedures devised for taking better photos, such as using the rule-of-thirds or foreground-background contrast. Note that the same term used for a factor might be interpreted very differently in each field. For instance, “contrast” in quality assessment may relate to levels adjustments that increase the range of the intensity values in the image, whereas for aesthetics, contrast may relate to the use of complementary colors to improve perceived object separation.

More formally, factors are ways in which an image can change, having the following properties:

1. a factor is identifiable, and its change is subjectively measurable; if the factor is identified, we say that it is present in an image
2. some factors can have a magnitude, such as blockiness artifacts due to JPEG compression, whereas others do not, such as framing, an aspect of composition
3. there is a level of change of a factor present in an image that causes the perceived rating of the image to change as well
4. most factors are not independent; they affect each other
5. factors have an associated level of complexity; some are simple, like sharpness, and some more complicated such as framing

The complexity of a factor (f_k^*) can be understood in terms of the complexity of a generative model H^* that reproduces all possible changes to an image in which only the specific factor changes. We denote this set of images as $D_{f_k^*}$, and we call it the span of f_k^* .

As most factors are not independent, it is difficult or impossible to define changes constrained to a single factor. However, if we were to limit ourselves to a small set of factors $F = \{f_1, f_2, \dots, f_n\}$, then we can talk about a change in factor f_k , that is significantly larger than the change in $F \setminus \{f_k\}$ given a subjective threshold for significance and a similar scale of the possible values that each factor can take. In this sense, we relax the constraint regarding changes to images that only affect a single factor but discuss changes that largely affect a particular factor and minimally impact other factors. Thus, having a model of a factor f_k in this new definition, its complexity is given by the minimum description length of the generative model H given the set of images D_{f_k} spanned by the factor f_k . This has been defined as $L(D_{f_k}, H)$ [19].

Such definitions provide us with a theoretical framework to consider the factors that influence perception, including technical and aesthetic quality. Quality itself can be viewed as a complex factor, affected by other simpler ones. It suggests approaches to systematically study quality subjectively:

1. for a particular type of quality, identify factors correlated with it
2. choose simpler factors (base factors) than the quality itself; more complex factors may just be related by having common causes with quality
3. determine their relationships to quality via subjective studies
4. select a diverse set of causal factors to control and validate further studies

The approach described above could enable more precise subjective quality assessment and provide annotations useful for the training of better multi-task-learning models. As technical and aesthetic quality are related, measuring their common factors would be beneficial to both. That might be a good starting point for future work. Next, we talk more about subjective measurements and why they are important for better quality assessment.

14.4.2 Psychometrics to the Rescue

Deep learning has helped make great strides in predicting aesthetics. However, in order to create better data-driven models, it is crucial to consider not only modeling techniques but subjective measurement methods as well.

Psychometric methods are concerned with the measurement of perceptual attributes. Their proper application is critical. Machine learning methods generally assume measurements are sufficiently accurate, which is not always true. Only by taking into account both measurement and modeling can we develop comprehensive and accurate aesthetic assessment methods.

As shown in previous sections, DNN IAA models are still lacking in their ability to predict accurately, trailing other related perceptual tasks such as IQA. Although DNNs have been shown to be noise-tolerant to some extent [12, 20, 37], they are also easily biased. If the aesthetic annotations are invalid, modeling that relies on them will be incorrect as well. However, in addition to measurement validity, content validity, or coverage of the target domain [18], is a requirement for developing better assessors. To ensure dataset quality in IAA and other perceptual tasks, previous research on IQA has proposed using ecological validity, as discussed in [26].

We argue that the under-performance of current IAA models, when compared to IQA models, stems in part from both required components: (1) algorithmic modeling and (2) quality of the data and annotations. Specifically, advancements in DNN architectures, training regimes, and losses should be incorporated into the IAA domain. With respect to the data, both the ecological validity of datasets and the reliability and validity of the subjective measurements ought to be improved. The former will impact the potential for generalization given powerful predictive

methods, while the latter ensures that the models learn to predict the correct quantities.

14.5 Conclusions

We have highlighted interconnections between aesthetic and technical quality assessment. By examining handcrafted features inspired by factors that are significant in assessing quality and aesthetics, we have shown commonalities between the two tasks. We saw that technical quality and aesthetics share perceptual features extracted from CNNs, helping us quantify the level of abstraction at which the overlap occurs. Moreover, we have shown that learning to predict one type of rating can positively impact the other, indicating that aesthetic and technical quality assessment are clearly not mutually exclusive but rather complementary.

Given the numerous connections between technical quality and aesthetics, as more of the problems in each field are addressed, we see a convergence of methods. Ideas that have resulted in improvements in IQA may shed light on why IAA performance is poor. Understanding and modeling IQA and IAA together may help improve both.

Thus, technical quality and aesthetics are blending together, and a more general form of visual quality assessment is emerging, or maybe quality is subsumed into aesthetics, which includes both IQA and IAA. This may start with more cross-task predictions and datasets annotated for both technical and aesthetic quality, but it seems that a general quality factor might be underlying both types of judgments.

As such, the integration of both approaches can lead to more robust and comprehensive evaluations of general quality, with potential applications in diverse fields such as art, design, marketing, and social media. Further research in this area has the potential to enhance our understanding of the complex relationships between perception, cognition, and technology, ultimately advancing our ability to create, evaluate, and appreciate visual media.

Acknowledgments Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 251654672–TRR 161, and Malaysia Ministry of Education FRGS Research Grant—Project-ID FRGS/1/2018/ICT02/MMU/02/2.

References

1. Amer, A., Dubois, E.: Fast and reliable structure-oriented video noise estimation. *IEEE Trans. Circuits Syst. Video Technol.* **15**(1), 113–118 (2005)
2. Babu, R.V., Suresh, S., Perkis, A.: No-reference JPEG-image quality assessment using GAP-RBF. *Signal Process.* **87**(6), 1493–1503 (2007)
3. Bianco, S., Celona, L., Napoletano, P., Schettini, R.: On the use of deep learning for blind image quality assessment. *Signal Image Video Process.* **12**(2), 355–362 (2018)

4. Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.* **27**(1), 206–219 (2017)
5. Cerosaletti, C.D., Loui, A.C.: Measuring the perceived aesthetic quality of photographic images. In: Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEX), pp. 47–52. IEEE (2009)
6. Chen, C., Bloom, J.A.: A blind reference-free blockiness measure. In: Pacific-Rim Conference on Multimedia, pp. 112–123. Springer (2010)
7. Chetouani, A., Beghdadi, A.: A new image quality estimation approach for jpeg2000 compressed images. In: Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 581–584. IEEE (2011)
8. Chetouani, A., Beghdadi, A., Chen, S., Mostafaoui, G.: A novel free reference image quality metric using neural network approach. In: Proceedings of the International Workshop on Video Processing Quality Metrics (VPMQ), pp. 1–4 (2010)
9. Ciancio, A., da Silva, E.A., Said, A., Samadani, R., Obrador, P., et al.: No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Trans. Image Process.* **20**(1), 64–75 (2010)
10. Cohen, E., Yitzhaky, Y.: No-reference assessment of blur and noise impacts on image quality. *Signal Image Video Process.* **4**(3), 289–302 (2010)
11. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 288–301. Springer (2006)
12. Denil, M., Shakibi, B., Dinh, L., Ranzato, M., De Freitas, N.: Predicting parameters in deep learning. Preprint. arXiv:1306.0543 (2013)
13. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3677–3686 (2020)
14. Farhat, F., Kamani, M.M., Wang, J.Z.: Captain: Comprehensive composition assistance for photo taking. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* **18**(1), 1–24 (2022)
15. Ferzli, R., Karam, L.J.: A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). *IEEE Trans. Image Process.* **18**(4), 717–728 (2009)
16. Gao, F., Yu, J., Zhu, S., Huang, Q., Tian, Q.: Blind image quality prediction by exploiting multi-level deep representations. *Pattern Recogn.* **81**, 432–442 (2018)
17. Ghazal, M., Amer, A.: Homogeneity localization using particle filters with application to noise estimation. *IEEE Trans. Image Process.* **20**(7), 1788–1796 (2010)
18. Götz-Hahn, F., Hosu, V., Lin, H., Saupe, D.: KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. *IEEE Access* **9**, 72139–72160. IEEE (2021)
19. Grünwald, P.D.: The Minimum Description Length Principle. MIT Press (2007)
20. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. Preprint. arXiv:1510.00149 (2015)
21. Hassen, R., Wang, Z., Salama, M.: No-reference image sharpness assessment based on local phase coherence measurement. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2434–2437. IEEE (2010)
22. 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 (CVPR), pp. 770–778 (2016)
23. He, S., Zhou, Z., Farhat, F., Wang, J.Z.: Discovering triangles in portraits for supporting photographic creation. *IEEE Trans. Multimedia* **20**(2), 496–508 (2017)
24. Hosu, V.: Beauty as amplified perception: Automatic artist-level light montage. PhD dissertation, National University of Singapore (2014)
25. Hosu, V., Goldlucke, B., Saupe, D.: Effective aesthetics prediction with multi-level spatially pooled features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9375–9383 (2019)

26. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.* **29**, 4041–4056 (2020)
27. Kahneman, D., Sibony, O., Sunstein, C.R.: *Noise: A Flaw in Human Judgment*. Hachette UK (2021)
28. Kandemir, B., Zhou, Z., Li, J., Wang, J.Z.: Beyond saliency: Assessing visual balance with high-level cues. In: Proceedings of the 1st International ACM Thematic Workshops, Thematic Workshops 2017, pp. 26–34 (2017)
29. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1733–1740 (2014)
30. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 419–426. IEEE (2006)
31. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5148–5157 (2021)
32. Kim, J., Lee, S.: Fully deep blind image quality predictor. *IEEE J. Sel. Top. Signal Process.* **11**(1), 206–220 (2016)
33. Kong, S., Shen, X., Lin, Z., Mech, R., Fowlkes, C.: Photo aesthetics ranking network with attributes and content adaptation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 662–679. Springer (2016)
34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012)
35. Kucer, M., Loui, A.C., Messinger, D.W.: Leveraging expert feature knowledge for predicting image aesthetics. *IEEE Trans. Image Process.* **27**(10), 5100–5112 (2018)
36. Lai, Y.K., Kuo, C.C.J.: A Haar wavelet approach to compressed image quality measurement. *J. Vis. Commun. Image Represent.* **11**(1), 17–40 (2000)
37. LeCun, Y., Denker, J.S., Solla, S.A., Howard, R.E., Jackel, L.D.: Optimal brain damage. In: NIPs, vol. 2, pp. 598–605. Citeseer (1989)
38. Leder, H., Hakala, J., Peltoketo, V.T., Valuch, C., Pelowski, M.: Swipes and saves: A taxonomy of factors influencing aesthetic assessments and perceived beauty of mobile phone photographs. *Front. Psychol.* **13**, 786977–786977 (2022)
39. Li, X.: Blind image quality assessment. In: Proceedings of the 9th IEEE International Conference on Image Processing (ICIP), vol. 1, pp. I–I. IEEE (2002)
40. Li, J., Yao, L., Wang, J.Z.: Photo composition feedback and enhancement: Exploiting spatial design categories and the notan dark-light principle. *Mobile Cloud Visual Media Computing: From Interaction to Service*, pp. 113–144 (2015)
41. Li, D., Jiang, T., Jiang, M.: Norm-in-norm loss with faster convergence and better performance for image quality assessment. In: Proceedings of the ACM International Conference on Multimedia, pp. 789–797 (2020)
42. Li, L., Zhu, H., Zhao, S., Ding, G., Lin, W.: Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Trans. Image Process.* **29**, 3898–3910 (2020)
43. Lin, K.Y., Wang, G.: Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 732–741 (2018)
44. Lin, W., Dong, L., Xue, P.: Visual distortion gauge based on discrimination of noticeable contrast changes. *IEEE Trans. Circ. Syst. Video Technol.* **15**(7), 900–909 (2005)
45. Lin, H., Hosu, V., Saupe, D.: DeepFL-IQA: Weak supervision for deep IQA feature learning. Preprint. arXiv:2001.08113 (2020)
46. Liu, H., Heynderickx, I.: A no-reference perceptual blockiness metric. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 865–868. IEEE (2008)

47. Liu, H., Heynderickx, I.: A perceptually relevant no-reference blockiness metric based on local image characteristics. *EURASIP J. Adv. Signal Process.* **2009**, 1–14 (2009)
48. Liu, H., Klomp, N., Heynderickx, I.: A no-reference metric for perceived ringing artifacts in images. *IEEE Trans. Circ. Syst. Video Technol.* **20**(4), 529–539 (2009)
49. Liu, H., Redi, J., Alers, H., Zunino, R., Heynderickx, I.: No-reference image quality assessment based on localized gradient statistics: application to JPEG and JPEG2000. In: *Human Vision and Electronic Imaging XV*, vol. 7527, pp. 419–427. SPIE (2010)
50. Liu, X., Tanaka, M., Okutomi, M.: Single-image noise level estimation for blind denoising. *IEEE Trans. Image Process.* **22**(12), 5226–5237 (2013)
51. Liu, X., Van De Weijer, J., Bagdanov, A.D.: Rankiq: Learning from rankings for no-reference image quality assessment. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1040–1049 (2017)
52. Liu, D., Puri, R., Kamath, N., Bhattacharya, S.: Composition-aware image aesthetics assessment. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (CVPR)*, pp. 3569–3578 (2020)
53. Lu, Z., Lin, W., Yang, X., Ong, E., Yao, S.: Modeling visual attention’s modulatory aftereffects on visual sensitivity and quality evaluation. *IEEE Trans. Image Process.* **14**(11), 1928–1942 (2005)
54. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rapid: Rating pictorial aesthetics using deep learning. In: *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 457–466 (2014)
55. Lu, X., Lin, Z., Shen, X., Mech, R., Wang, J.Z.: Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 990–998 (2015)
56. Luo, Y., Tang, X.: Photo and video quality evaluation: Focusing on the subject. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 386–399. Springer (2008)
57. Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., Zuo, W.: End-to-end blind image quality assessment using deep neural networks. *IEEE Trans. Image Process.* **27**(3), 1202–1213 (2017)
58. Mai, L., Jin, H., Liu, F.: Composition-preserving deep photo aesthetics assessment. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 497–506 (2016)
59. Marichal, X., Ma, W.Y., Zhang, H.: Blur determination in the compressed domain using DCT information. In: *Proceedings of the International Conference on Image Processing*, vol. 2, pp. 386–390. IEEE (1999)
60. Marziliano, P., Dufaux, F., Winkler, S., Ebrahimi, T.: Perceptual blur and ringing metrics: application to jpeg2000. *Signal Process. Image Commun.* **19**(2), 163–172 (2004)
61. Men, H., Lin, H., Saupe, D.: Spatiotemporal feature combination model for no-reference video quality assessment. In: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3 (2018)
62. Murray, N., Gordo, A.: A deep architecture for unified aesthetic prediction. Preprint. arXiv:1708.04890 (2017)
63. Murray, N., Marchesotti, L., Perronnin, F.: AVA: A large-scale database for aesthetic visual analysis. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2408–2415. IEEE (2012)
64. Ouni, S., Zagrouba, E., Chambah, M., Herbin, M.: No-reference image semantic quality approach using neural network. In: *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 106–113. IEEE (2011)
65. Pan, D., Shi, P., Hou, M., Ying, Z., Fu, S., Zhang, Y.: Blind predicting similar quality map for image quality assessment. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6373–6382 (2018)
66. Peli, E.: Contrast in complex images. *JOSA A* **7**(10), 2032–2040 (1990)
67. Pyatykh, S., Hesser, J., Zheng, L.: Image noise level estimation by principal component analysis. *IEEE Trans. Image Process.* **22**(2), 687–699 (2012)

68. Rank, K., Lendl, M., Unbehauen, R.: Estimation of image noise variance. *IEE Proc. Vis. Image Signal Process.* **146**(2), 80–84 (1999)
69. Redi, J.A., Heynderickx, I.: Image integrity and aesthetics: towards a more encompassing definition of visual quality. In: *Human Vision and Electronic Imaging XVII*, vol. 8291, p. 829115. International Society for Optics and Photonics (2012)
70. Schubert, L., Petrov, M., Carter, S., Cammarata, N., Goh, G., Olah, C.: OpenAI microscope (2020). <https://openai.com/blog/microscope/>
71. Shahid, M., Rossholm, A., Lövström, B., Zepernick, H.J.: No-reference image and video quality assessment: a classification and review of recent approaches. *EURASIP J. Image Video Process.* **2014**(1), 1–32 (2014)
72. Sheng, K., Dong, W., Ma, C., Mei, X., Huang, F., Hu, B.G.: Attention-based multi-patch aggregation for image aesthetic assessment. In: *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 879–886 (2018)
73. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Preprint. arXiv:1409.1556 (2014)
74. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., Zhang, Y.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3667–3676 (2020)
75. Su, S., Hosu, V., Lin, H., Zhang, Y., Saupe, D.: Koniq++: Boosting no-reference image quality assessment in the wild by jointly predicting image quality and defects. In: *Proceedings of the 32nd British Machine Vision Conference (BMVC)*, vol. 2 (2021)
76. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31 (2017)
77. Talebi, H., Milanfar, P.: NIMA: Neural image assessment. *IEEE Trans. Image Process.* **27**(8), 3998–4011 (2018)
78. Tinio, P.P., Leder, H.: Natural scenes are indeed preferred, but image quality might have the last word. *Psychol. Aesthet. Creat. Arts* **3**(1), 52 (2009)
79. Tinio, P.P., Leder, H., Strasser, M.: Image quality and the aesthetic judgment of photographs: Contrast, sharpness, and grain teased apart and put together. *Psychol. Aesthet. Creat. Arts* **5**(2), 165 (2011)
80. Tong, H., Li, M., Zhang, H., Zhang, C.: Blur detection for digital images using wavelet transform. In: *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, vol. 1, pp. 17–20. IEEE (2004)
81. Tong, H., Li, M., Zhang, H.J., He, J., Zhang, C.: Classification of digital photos taken by photographers or home users. In: *Pacific-Rim Conference on Multimedia*, pp. 198–205. Springer (2004)
82. Winkler, S., Vandergheynst, P.: Computing isotropic local contrast from oriented pyramid decompositions. In: *Proceedings of the 6th IEEE International Conference on Image Processing (ICIP)*, vol. 4, pp. 420–424. IEEE (1999)
83. Wong, L.K., Low, K.L.: Saliency-enhanced image aesthetics class prediction. In: *Proceedings of the 16th IEEE International Conference on Image Processing (ICIP)*, pp. 997–1000. IEEE (2009)
84. Wu, J., Zeng, J., Liu, Y., Shi, G., Lin, W.: Hierarchical feature degradation based blind image quality assessment. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 510–517 (2017)
85. Yan, B., Bare, B., Tan, W.: Naturalness-aware deep no-reference image quality assessment. *IEEE Trans. Multimedia* **21**(10), 2603–2615 (2019)
86. Yang, Y., Xu, L., Li, L., Qie, N., Li, Y., Zhang, P., Guo, Y.: Personalized image aesthetics assessment with rich attributes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19861–19869 (2022)
87. Yao, L., Suryanarayanan, P., Qiao, M., Wang, J.Z., Li, J.: Oscar: On-site composition and aesthetics feedback through exemplars for photographers. *Int. J. Comput. Vis.* **96**, 353–383 (2012)

88. Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram, D., Bovik, A.: From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3575–3585 (2020)
89. Zeng, H., Cao, Z., Zhang, L., Bovik, A.C.: A unified probabilistic formulation of image aesthetic assessment. *IEEE Trans. Image Process.* **29**, 1548–1561 (2019)
90. Zhai, G., Zhang, W., Yang, X., Lin, W., Xu, Y.: No-reference noticeable blockiness estimation in images. *Signal Process. Image Commun.* **23**(6), 417–432 (2008)
91. Zhang, B., Niu, L., Zhang, L.: Image composition assessment with saliency-augmented multi-pattern pooling. Preprint. arXiv:2104.03133 (2021)
92. Zhu, X., Milanfar, P.: A no-reference sharpness metric sensitive to blur and noise. In: Proceedings of the IEEE International Workshop on Quality of Multimedia Experience (QoMEX), pp. 64–69. IEEE (2009)
93. Zhu, X., Milanfar, P.: A no-reference image content metric and its application to denoising. In: Proceedings of the 17th IEEE International Conference on Image Processing (ICIP), pp. 1145–1148. IEEE (2010)
94. Zhu, H., Li, L., Wu, J., Zhao, S., Ding, G., Shi, G.: Personalized image aesthetics assessment via meta-learning with bilevel gradient optimization. *IEEE Trans. Cybern.* **52**, 798 (2020)
95. Zuo, B.X., Tian, J.W., Ming, D.L.: A no-reference ringing metrics for images deconvolution. In: Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition, vol. 1, pp. 96–101. IEEE (2008)

Chapter 15

Image Restoration for Beautification



Dejia Xu, Yifan Jiang, and Zhangyang Wang

Abstract Image restoration refers to the recovery from a degraded image to achieve better visual aesthetics. With the advances in imaging and computing, image restoration has developed rapidly, dealing with the degradations occurring during image formation, transmission, and storage. It covers a wide range of topics, including image denoising, image deblurring, image inpainting, image super-resolution, reflection removal, light enhancement, etc. This chapter will cover some of the latest research results related to image restoration for beautification. For young minds entering this field, it will include surveys and tutorials on some foundational topics. And for veteran researchers, it will serve as a quick reference.

15.1 Introduction

Photo/V-log capturing is becoming a staple in modern urban life. Despite advancements in imaging and computing technology, images still turn out subpar in certain circumstances, like when capturing moving subjects, in low-light conditions, or during inclement weather. The drive to enhance image quality for the sake of human perception has spurred rapid growth in the field of image restoration.

Research on image restoration for beautification aims at improving image quality to fascinate human beings. The target inherently differs from the line of work that enhances the images for machine intelligence relying on computer vision algorithms such as night-time autonomous driving and biometric recognition.

Classic methods formulate the image restoration task as an inverse problem, where a degraded image is usually modeled as the result of matrix multiplication. Some early approaches assume the degradation matrix is known and adopt deconvolution [11, 62, 86]. In many other situations, the degradation matrix is unavailable, and additional regularizations [60, 82] are utilized to handle the ill-posed tasks.

D. Xu (✉) · Y. Jiang · Z. Wang

Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA

e-mail: dejia@utexas.edu; yifanjiang97@utexas.edu; atlaswang@utexas.edu

The field of image restoration also benefits from the recent advancements in deep learning. However, they generally perform well in certain domains and can not generalize well to complicated real-world cases. In this paper, we will provide a thorough survey of deep learning-based methods, especially for their superior performance.

Innovations in neural network architecture for image restoration have been a hot topic in the research community, with various proposals being put forth and adopted. Popular principles such as residual learning, multi-branch, and multi-scale learning have become widely used. Several network blocks have been developed to restore images effectively, with the main trends being the integration of frequency domain information, color space conversion, efficient convolution, dynamic blocks, attention mechanisms, and transformer blocks. Another important direction is the adoption of generative priors, as generative networks are capable of creating photo-realistic images without artifacts.

Apart from the development of network architecture, different loss terms have also been studied extensively. From early pixel-wise loss, such as L1, L2, and Charbonnier loss [6] to perceptual loss [28] and color loss [77] that improve different aspects of the visual quality, the choice of the most suitable loss functions is always in debate, and we are always in need of a better loss function to optimize the network. There are other attempts to utilize adversarial loss [36] for improving high-frequency details and image quality assessment (IQA) based loss, such as SSIM loss [75] and LPIPS loss [106], to optimize the metrics directly.

Additionally, the way we evaluate these methods is also evolving. Initially, the datasets were limited in size, but as time passed, more and more data was collected to support the training of large data-driven methods. Furthermore, the quality of the evaluation sets has been improved to ensure comprehensive comparisons. Early datasets consisted of synthetic paired images generated using a kernel (e.g. bicubic, gaussian blur) due to the limited availability of real-world paired data. Later, GANs were widely utilized to enhance the results of physics-based degradation models and provide more realistic synthetic data [66]. More recently, real-world paired datasets have been constructed by incorporating RAW images [7] and video information [78] with optical-flow-based alignment.

Additionally, evaluation metrics are important factors to discuss for researchers. There are a variety of different quality assessment metrics as convenient as PSNR, SSIM, and LPIPS. But these metrics do not align very well with the human visual system. MOS, on the other hand, represents the human visual system better, but requires a lot of work to collect enough samples to deliver an unbiased estimate. The trade-off of efficiency and quality remains an issue worth investigating for designing image restoration IQA metrics.

The remainder of this chapter is organized as follows. Section 15.1 introduces a wide variety of tasks involved in the field of image restoration. Section 15.3 presents the main trends in method design, including the model framework and learning strategies. Section 15.4 gives a comprehensive introduction to evaluation protocols, including the datasets and the evaluation metrics. Section 15.5 discusses

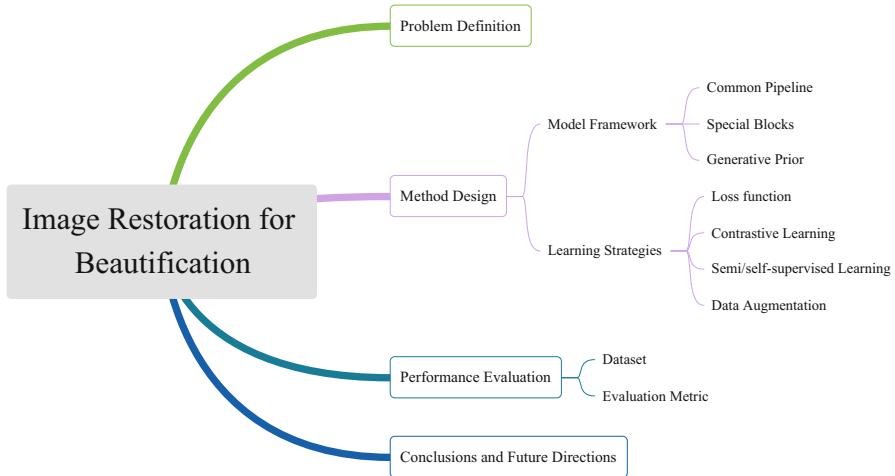


Fig. 15.1 Overview of the chapter

potential directions and challenges and draws a conclusion. An overview is provided in Fig. 15.1.

15.2 Problem Definition

Despite the rapid development of imaging technique, capturing images in real-world cases suffer from various kinds of degradations. In this section, we will discuss several common degradations and the corresponding image restoration task. A visual illustration of these tasks is provided in Fig. 15.2.

Image denoising focuses on removing noise from images. Although experts have been trying to improve imaging quality for decades, noises such as read-out noise are inevitable during capture. Additionally, noise is amplified in unpleasant lighting conditions, such as at night. The annoying noise is also still severe when capturing images using low-end cameras.

Image deblurring refers to recovering a clear image from obscure observation. Blur in the image is largely related to shutter speed, focus, and motion. The faster the shutter speed is, the less likely blur will exist. However, in special cases, such as low-light conditions, the shutter speed is expected to be low to capture enough photons to maintain a good enough brightness of the overall image. This leads to a trade-off between brightness and blurriness. In addition, blur exists in regions that are out of focus. If the object is not on the focal plane, then it is very likely to be blurry because the light rays do not converge. In addition, both the camera and the captured object usually move during the capture period. Such motion usually leads to a blurry trajectory in the final captured image.

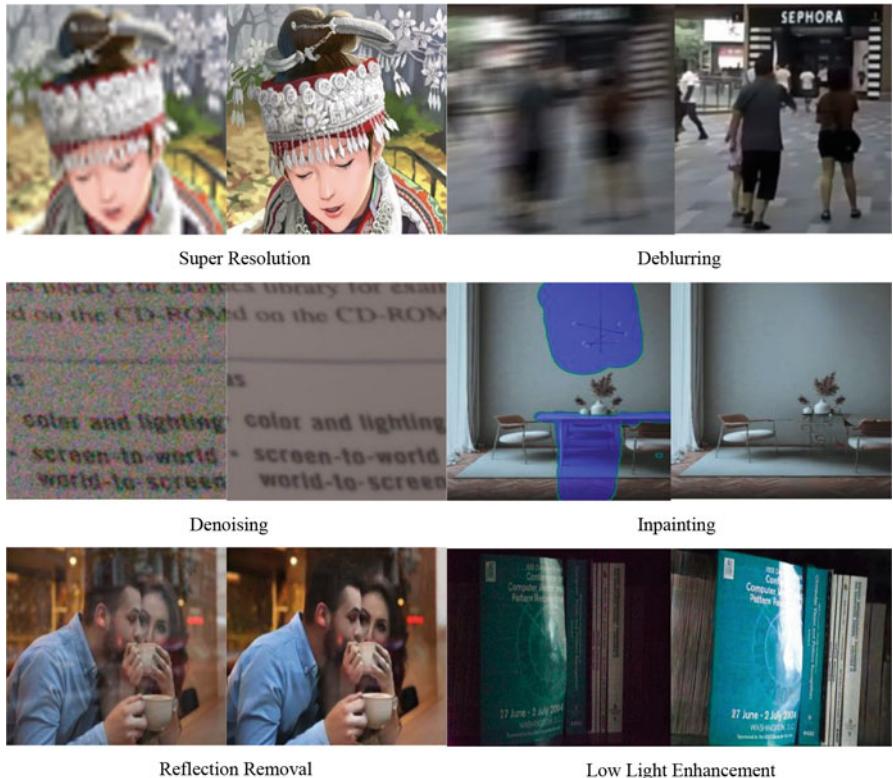


Fig. 15.2 Visual examples of various tasks in image restoration for beautification

Image inpainting refers to editing an image by removing and refilling pixels in a certain region. As a widely used application in image manipulation, image inpainting enhances human's ability to control the visual contents by re-designing the undesired regions. For example, when taking a group photo, pedestrians sometimes walk into the scene, and inpainting techniques are used to remove unexpected pedestrians.

Image super-resolution focuses on enlarging the spatial resolution of images. People usually want to zoom into an existing image and see more details inside a certain region. In such cases, image super-resolution methods transform the image into a new one with a bigger spatial resolution.

Reflection in images is introduced by reflective surfaces such as glass, mirrors, and windows. These objects are common in daily life, and we already get used to the reflections in our visual system. However, there are also times when such reflection is undesired. For example, when taking a photo in a museum, the protecting glass will involve reflection containing undesired contents that hurt the visual-looking experience of the images.

Light enhancement restores various kinds of images taken in the undesired light condition. For example, in low light conditions, noise and blur are amplified during capture, calling for the great need to enhance the visual looking. Additionally, low light leads to a darkened scene, and it is necessary to improve the scene's brightness.

There are also other image restoration tasks and many attempts to handle them, such as image demoiréing [18, 19, 91, 103, 104], deblocking [107, 108], and restoration from severe weather conditions (rain [95, 96], fog [38, 115], snow [48], etc.) Their relation with beautification is less close, so we refer the reader to the above-mentioned papers.

Recently, there has been a trend to handle multiple degradations together because, in real-world cases, images usually suffer from a complex degradation that is hard to model explicitly.

15.3 Method Design

Image restoration has been a long-standing task for decades and has attracted thousands of researchers pushing the performance boundary. The community has investigated various solutions to enhance the images for beautification. In this section, we will give a brief introduction to early classic methods and then summarize the main trends in learning-based methods in various aspects, including the model design and learning strategy.

15.3.1 *Model Framework*

15.3.1.1 Common Pipeline

Numerous network designs have been proposed with the rapid development of deep learning. The improvement of image classification architecture also benefits downstream tasks in image restoration. There are a few common pipelines or essential strategies for network design, and we introduce them one by one in the following paragraphs.

Residual Learning

The success of ResNet has inspired the idea of residual learning in network architecture design. It is widely observed that the global residual learning proved is beneficial for image restoration [69, 91] since it enables the network to only focus on generating the residual map instead of learning to generate the whole output directly. This design eases the network learning burden since the residual map is

considered to be simpler than the overall image. Furthermore, the idea of residual learning has found its way into the design of basic blocks [32, 91, 109].

Multi-Branch Learning

Multi-branch learning usually passes intermediate features on different paths with multiple operations before merging them into a single branch. Such design enables the decomposition of input features and naturally supports the implementation of the Retinex Theory, where images are decomposed into a reflection layer and an illumination layer. It is also widely adopted to use a multi-branch network to process the low-frequency component and high-frequency component separately or learn to process the intensity as well as the color information separately. Similarly, the idea of multi-branch is also successful in designing effective network blocks [39, 68, 92].

Multi-Scale Learning

The success of feature pyramid network (FPN) [44] and U-Net [63] in high-level tasks inspires the design of image restoration networks. The adoption of multi-scale learning has gained tremendous popularity since then. Resizing the images to multi-scale enables the network to restore at different scales. For example, a large object after downsampling will become small enough to be captured by the network's effective receptive field. The input image can be resized not only to multiple resolutions [96], but the intermediate feature can also be interpolated at different spatial sizes [13, 34]. Since the early layers of neural networks extract low-level features and deep layers extract high-level features, the utilization of multi-scale features is seen as a way to combine multiple frequency information and let the network learn the feature relationships across scales.

15.3.1.2 Special Blocks

Frequency Domain

While most work operates in the spatial domain of images, there are a few works investigating frequency domain processing. The human visual system is good at perceiving low-frequency changes in images and is not that accurate at capturing fine detail and high-frequency components. Image restoration networks, as a result, propose to handle different frequency components separately in order to perform an adaptive enhancement. DCT and Wavelet are two widely used transformations to transform images between the spatial domain and the frequency domain.

Color Space

Most existing work operates on sRGB color space directly because it is the most commonly used color space for phones and cameras. There are a few works that operate on the YUV or YCbCr color space and process Y and CbCr separately. Since the Y channel consists of intensity information and the CbCr channel dominates the color information, processing them separately can ease the network burden and help reduce color-shifting artifacts in results. There are also numerous attempts to directly process raw RGB. Raw RGB contains a high dynamic range and is not processed by the ISP, so directly operating on raw RGB usually leads to much better performance compared to sRGB color space [7]. Usually referred to as neural ISP, these works take raw RGB images as input and output a restored sRGB image.

Efficient Convolution

The convolution layer has been widely used in deep neural networks. In order to speed up the neural network, a direct idea is to improve this basic building block. For example, dilated convolution [83, 99] focuses on enlarging the receptive field and introducing contextual information to achieve better performance for image super resolution [110] and low light enhancement [61]. Group convolution [33] and depth-wise convolutions [10] are essentially useful for lightweight CNN design and can largely reduce the number of parameters and FLOPs for a variety of tasks [34, 42, 59].

Dynamic Block

Apart from the above-mentioned neural network design with one-for-all parameters, there is another line of work utilizing adaptive weights. Generally, dynamic blocks consist of two branches, one of which generates the weight of the other branch. In this way, the convolution parameters are adapted for the input features and can adequately process the features instead of applying a single set of kernels to the entire input feature. Adaptive instance normalization (AdaIN) [24] was originally proposed for style transfer. It learns the coefficients to perform adaptive affine transformations on features and has been found to be useful for image harmonization [47], denoising [31], and deblurring [3]. Dynamic convolutions [45, 101, 112] and kernel-prediction networks [5, 52] predict dynamic filters for each image. HDRNet [16] and MalleConv [27] further speed up the process by processing on a downsampled image or feature map and later performing slicing to obtain the final result.

Attention Mechanism

Attention mechanism has been favored by researchers since it resembles the human visual system to find significance in complex scenes [17]. One line of work adopts channel attention [12, 21, 46], which learns to assign a weight for each channel to emphasize the important features. Another line of work uses spatial attention [50, 56, 76, 87, 114], where a spatial weight is adaptively learned and helps the network focus on more important regions.

Transformer Block

Recently, non-convolutional layers are also gaining popularity as vision transformers demonstrate their superiority in image classification. Self-attention [73, 111] calculates pairwise similarity and is capable of capturing long-range dependency. Recently, vision transformers [51, 72, 94] have also had great success in image restoration.

15.3.1.3 Generative Prior

After the success of StyleGAN in realistic image synthesis, StyleGANs have been found beneficial to image restoration tasks as well. With the help of GAN inversion techniques [90], one can map an in-the-wild image into the latent space of StyleGAN. Then, the latent is feed-forwarded into the StyleGAN to produce a new image of similar visual appearance but much better quality. GFPGAN [80] and GPEN [97], for example, adopt an encoder-based GAN inversion module to map the real-world degraded face images into the StyleGAN latent space. The StyleGAN's ability to generate high-quality faces is well observed and can be used to effectively restore low-quality faces in the wild. MyStyle [55] goes one step further to fine-tune StyleGAN using a couple of reference images and is capable of generating identity-specific images.

15.3.2 Learning Strategies

15.3.2.1 Loss Function

Apart from the attempts to improve the expressive ability of the neural network, many efforts have been made to design a better loss function for faster and better convergence.

Pixel-Wise Loss

Early methods use \mathcal{L}_2 loss to regress the desired images in a fully supervised setting. Later in [113], researchers discovered that the \mathcal{L}_2 loss tends to generate a blurry result compared to using \mathcal{L}_1 loss. Charbonnier loss [6] introduces robustness with a continuous parameter and further improves the image restoration tasks. It is defined as follows,

$$\mathcal{L}(\hat{I}, I) = \frac{1}{hwc} \sum_{i,j,k} \sqrt{\left(\hat{I}_{i,j,k} - I_{i,j,k}\right)^2 + \epsilon^2} \quad (15.1)$$

Apart from these losses where each pixel is treated equally, there are various attempts to emphasize specific spatial regions, such as using edge maps [65] and semantic maps [14].

Adversarial Loss

GANs are also involved in improving the quality further. SRGAN [36] discovered that adversarial loss improves the perceptual quality of the network output compared to the \mathcal{L}_1 loss. An additional discriminator is adopted to distinguish whether an image is a super-resolved result from the generator network or it is indeed a natural image. Using adversarial loss can empirically improve the textures of the generated images. It is widely believed that the discriminator network can push the output of the network towards the natural image manifold. With the development of better GAN techniques, there are numerous adversarial learning attempts to improve image quality [57, 71].

Perceptual Loss

Perceptual loss [28] is widely used to measure the semantic distance between images. With the help of a pre-trained image classification network, the perceptual loss is defined as the euclidean distance between two high-level feature maps of two images. Pre-trained on ImageNet, the classification network transfers knowledge of image semantics to the image restoration network. There are also attempts to build discriminators based on deep network features. This combination of perceptual loss and adversarial loss has been proven to be successful in the Perceptual Discriminator [67].

Color Loss

There are numerous attempts to regularize color distortion in image restoration tasks. One line of work operates on YUV color space. By processing the Y channel and UV channel separately, the network can individually learn to restore the intensity as well as the desired color.

Another line of work focuses on DeepUPE [77] formulates the RGB color as a three-dimensional vector, and calculates the angular distance between the output and the ground truth,

$$\mathcal{L} = \sum_p \angle \left((\mathcal{F}(I_i))_p, (\tilde{I}_i)_p \right), \quad (15.2)$$

where $(\cdot)_p$ denotes a pixel and $\angle(\cdot, \cdot)$ refers to the operator that calculates the angle distance.

Prior-Based Loss

There are a lot of hand-crafted priors that have been shown successful in image restoration. For example, in [2, 36, 49, 102], researchers use total variation (TV) loss to enforce the spatial smoothness of images, as follows,

$$\mathcal{L}_{\text{TV}}(\hat{I}) = \frac{1}{hwc} \sum_{i,j,k} \sqrt{\left(\hat{I}_{i,j+1,k} - \hat{I}_{i,j,k} \right)^2 + \left(\hat{I}_{i+1,j,k} - \hat{I}_{i,j,k} \right)^2}. \quad (15.3)$$

On the other hand, when processing images of specific domains, such as the face, human body, etc. Numerous works are utilizing face landmarks [40] or 3d prior [23] to regularize undesired artifacts in the network's output.

IQA-Based Loss

Another popular design of the loss function is to directly optimize the image quality assessment (IQA) metric. When the interested IQA is differentiable, the loss function will be the IQA itself. For example, SSIM [75], NIQE [54], BRISQUE [53] and LPIPS [106] have been proven successful in many image restoration tasks. However, when the target IQA is not differentiable, a special design is adopted such as deep reinforcement learning [22, 79, 100] is adopted.

Cycle Consistency Loss

The idea of cycle consistency is first introduced to the computer vision community in CycleGAN [116]. It has also been proven successful in image restoration in order to help the restoration network learn a lossless representation. Not only is there a network learning to restore the image, but another network is also utilized to degrade the images as well. The cycle consistency enforces the network to learn rich representation with no information loss. The overall loss function is formulated as follows,

$$\mathcal{L} = \|\text{Degrade}(\text{Restore}(x)) - x\|. \quad (15.4)$$

Such design is also orthogonal to the above-mentioned loss functions that minimize the distance between two images and can be used together to improve the performance further. In practice, researchers usually simultaneously use multiple losses to constrain different aspects of the network output.

15.3.2.2 Contrastive Learning

Contrastive Learning has demonstrated success in self-supervised representation learning [25, 35] and has produced comparable results to the state-of-the-art supervised methods on the ImageNet dataset. The idea of contrastive learning is to push original and augmented images closer, since they have similar contents, and to push original and negative images away, since they have different contents. It also shows excellent improvement in various image restoration tasks [9, 43, 88, 89]. Specifically, in the restoration context, the degraded images are usually considered to be instance-level augmentations to generate positive samples. The negative samples are obtained from other images inside the mini-batch. More recently, in [9, 58], the patch-level similarity is also adopted to enforce better representation learning.

15.3.2.3 Semi-supervised Learning and Self-supervised Learning

Since the inevitable difficulty in collecting real-world paired images for large-scale training, the ubiquitous need for unpaired training has attracted numerous researchers. Existing semi-supervised and self-supervised learning methods for image classification, such as pseudo-label [37], transfer learning [84], and consistency regularization [96] are widely adopted. On the other hand, there are also attempts [41] using various priors, such as total variation loss, to regularize the unlabeled set.

15.3.2.4 Data Augmentation

Less attention has been made to building a data augmentation mechanism specialized for image restoration. For decades, a popular data augmentation strategy has been to use flipping, transposing, and cropping, which are commonly used for image classification. More recently, CutBlur [98] presents an effective data augmentation for image super-resolution.

On the other hand, test-time enhancement is widely used to further enhance image restoration performance, especially for challenges such as NTIRE [104] and AIM [85]. Also known as self-ensemble, it refers to the technique of performing several data augmentations to the input images at test time and then merging the results using mean or summation. As shown in AFN [91] and HINet [8], the self-ensemble mechanism can further improve the performance of the model by a large margin without the need to perform additional training tricks.

15.4 Performance Evaluation

With the rapid development of image restoration techniques, more attention has been given to evaluating the performance of existing methods. In this section, we will first briefly introduce the dataset being used and then discuss the metrics for comparison.

15.4.1 Dataset

15.4.1.1 Synthetic Data

For the learning-based method, there is a heavy need for large-scale datasets that facilitate the training of neural networks. For many image restoration tasks, however, it is essentially hard to capture well-aligned real-world paired data. The inevitable misalignment issue comes mainly from the movement of objects in daily life. It is hard to find diverse static scenes in practice.

To overcome this situation, early approaches choose to adopt synthetic images for training. Usually, a task-specific degradation model is obtained and applied to natural images in high-quality image datasets, such as BSD500, ImageNet, ADE20k.

However, these datasets are not specially designed for image restoration tasks. As a result, their scene diversity and image quality are not desirable. Agustsson et al. [1] and Karras et al. [29] are proposed to facilitate network training on high-quality datasets.

The hand-crafted models are far from perfect for covering diverse real-world scenarios. Researchers also attempt to use a neural network to learn the degradation

directly. After generative adversarial network (GAN) was introduced to the community, it has been widely applied to refine the results generated from physical-based models. More recently, rendering software is also used to collect data.

There are also several attempts to alleviate the need for large amounts of paired images. For example, with the help of GANs, EnlightenGAN [26] does not require paired data during training. Unsupervised training and semi-supervised training [93] are also widely explored among the community.

15.4.1.2 Real Data

Several researchers have focused their efforts on collecting real-world data. By referring to real-world data, we are referring to the dataset where the low-quality images are also from the real world. One line of work collects the limited paired data as the test set. The performance gap between synthetic datasets and real datasets demonstrates the importance of using real data for evaluation. Other works utilize additional alignment techniques to reduce the effect of real-world undesired motion. For example, [4] uses optical flow to warp the frames. [7] uses RAW image and [78] uses the video restoration technique to generate better results as ground truth to construct a paired image dataset.

15.4.2 Evaluation Metric

Despite the significant progress in image restoration techniques, the community remains unsure of how to perform a quality assessment for images. Image quality generally refers to the perceptual assessment of human beings. As highly subjective as it is, researchers have investigated numerous resources in search of better metrics that align better with human perceptions. Next, we'll introduce several widely used IQA methods for image restoration.

Classics Methods

One of the most popular quality measurements for image restoration is the Peak signal-to-noise ratio (PSNR). We only need the maximum pixel value and the mean squared error to calculate the PSNR. Since it measures the pixel-wise difference between images, directly optimizing it usually leads to poor overall visual quality. The structural similarity index (SSIM) [75], on the other hand, is widely used due to its similarity to the human visual system in extracting structures from images [74]. SSIM compares the luminance, contrast, and structures of images. There are also other IQA methods, such as the Natural Image Quality Evaluator (NIQE) [54], the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [53], and the feature similarity (FSIM) [105].

Apart from attempts at overall fidelity in image quality, there are also task-specific evaluation metrics. For example, Angular error [20] measures the color distortion.

Learning-Based Metrics

Quality assessment methods also benefit from the rapid development of deep learning. With the help of a large dataset, researchers are able to train neural networks that predict the image perceptual quality. NIMA [70] predicts quality scores directly. DeepQA [30] predicts visual similarity between images. LPIPS [106] uses the difference in deep network features to show the perceptual difference between images. Although these methods demonstrate superior performance in aligning with the human visual system, there is still a gap between human evaluation and existing metrics.

Human Evaluation

Although there are various attempts to propose metrics for image quality assessment, the high subjectiveness of image quality calls for the need for human evaluation. Mean opinion score (MOS) is a commonly used metric to quantify human rating. Human raters are asked to give scores in a certain range to images, and the MOS is defined to be the mean of the scores. Although MOS seems to be a perfect metric, it still suffers from assessment bias in practice because of the limited raters.

15.5 Conclusions and Future Directions

In this survey, we present an extensive summary of recent advances in image restoration for beautification. We discuss the various task definitions, model designs, and evaluation protocols thoroughly. Despite the exciting progress, there are still a few problems remaining to be solved. In the following paragraphs, we will summarize some major challenges and possible directions to improve.

Efficiency

Existing image restoration networks are mostly heavy to obtain high capacity and perform well. However, in most use cases, neural networks are expected to run inferences on edge devices, such as mobile phones. Under such circumstances, hardware restrictions lead to a smaller network design space. As a result, there is a strong need to design efficient image restoration networks. Apart from hand-crafting network architectures, recent progress in neural architectural search (NAS) is also promising for hardware-friendly network design.

Real-World Image Restoration

Existing works that meet academic benchmarks do not always generalize to real-world testing cases, as shown in recent real-world super-resolution works [81]. The challenge of dealing with real-world test cases is that the data is usually contaminated with multiple degradations, such as blur, noise, and JPEG com-

pression. Furthermore, in the context of real-world cases, we need to perform blind restoration, which means we do not know how the images are degraded. Handling these challenges requires the network to be robust against different kinds of degradations, while in the meantime, shouldn't generate artifacts that deteriorate the visual looking of images. Artifacts are hard to prevent when inferring from diverse real-world images, hindering image restoration techniques from being widely applied in production.

Quality Assessment

Although there exist various quality assessment methods, there are still gaps between the human visual system and quality metrics. Recently there have been works on improving the quality of the metrics, and there is a trade-off between efficiency and accuracy. PSNR, SSIM, and LPIPS are widely used and efficient, but they do not align well enough with the human visual system. MOS, on the other hand, is accurate enough to represent the human visual system, but to obtain an unbiased result, one needs to put in a lot of effort. Additionally, for real-world testing cases especially, there are only a few blind IQA attempts that exist, and they do not generalize well to the diverse combination of degradations in real life.

Aesthetic-Driven Image Restoration

Although the purpose of most restoration techniques is to beautify the photo, not all restoration operations create more beautiful photos. For example, some blurry or lighting effects might be deliberately shot to satisfy aesthetic purposes. Especially for portrait mode and pictorial images, how to identify intentional artifacts and avoid distorting them remains an open question. Recently, LAION-Aesthetics Image Dataset [64] is collected with aesthetic ratings of images, and aesthetic gradient [15] is proposed to drive text-to-image generation towards the aesthetics of the user. These approaches might shed light on a promising direction toward the goal of aesthetic-driven image restoration.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126–135 (2017)
2. Aly, H.A., Dubois, E.: Image up-sampling using total-variation regularization with a new observation model. *IEEE Trans. Image Process.* **14**(10), 1647–1659 (2005)
3. An, S., Roh, H., Kang, M.: Blur invariant kernel-adaptive network for single image blind deblurring. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2021)
4. Ba, Y., Zhang, H., Yang, E., Suzuki, A., Pfahl, A., Chandrappa, C.C., de Melo, C., You, S., Soatto, S., Wong, A., et al.: Towards ground truth for single image deraining. Preprint. arXiv:2206.10779 (2022)
5. Bako, S., Vogels, T., McWilliams, B., Meyer, M., Novák, J., Harvill, A., Sen, P., Derose, T., Rousselle, F.: Kernel-predicting convolutional networks for denoising Monte Carlo renderings. *ACM Trans. Graph.* **36**(4), 97–1 (2017)

6. Barron, J.T.: A general and adaptive robust loss function. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4331–4339 (2019)
7. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3291–3300 (2018)
8. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: Hinet: Half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 182–192 (2021)
9. Chen, X., Pan, J., Jiang, K., Li, Y., Huang, Y., Kong, C., Dai, L., Fan, Z.: Unpaired deep image deraining using dual contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2017–2026 (2022)
10. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
11. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans. Image Process.* **16**(8), 2080–2095 (2007)
12. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11065–11074 (2019)
13. Fan, Z., Wu, H., Fu, X., Huang, Y., Ding, X.: Residual-guide network for single image deraining. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 1751–1759 (2018)
14. Fan, M., Wang, W., Yang, W., Liu, J.: Integrating semantic segmentation and retinex model for low-light image enhancement. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2317–2325 (2020)
15. Gallego, V.: Personalizing text-to-image generation via aesthetic gradients. Preprint. arXiv:2209.12330 (2022)
16. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. *ACM Trans. Graph. (TOG)* **36**(4), 1–12 (2017)
17. Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M.: Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **8**(3), 331–368 (2022)
18. He, B., Wang, C., Shi, B., Duan, L.Y.: Mop moire patterns using mopnet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2424–2432 (2019)
19. He, B., Wang, C., Shi, B., Duan, L.Y.: Fhde 2 net: Full high definition demoiréing network. In: European Conference on Computer Vision, pp. 713–729. Springer (2020)
20. Hordley, S.D., Finlayson, G.D.: Re-evaluating colour constancy algorithms. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, vol. 1, pp. 76–79. IEEE (2004)
21. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
22. Hu, Y., He, H., Xu, C., Wang, B., Lin, S.: Exposure: A white-box photo post-processing framework. *ACM Trans. Graph. (TOG)* **37**(2), 1–17 (2018)
23. Hu, X., Ren, W., LaMaster, J., Cao, X., Li, X., Li, Z., Menze, B., Liu, W.: Face super-resolution guided by 3d facial priors. In: European Conference on Computer Vision, pp. 763–780. Springer (2020)
24. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
25. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. *Technologies* **9**(1), 2 (2020)
26. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.* **30**, 2340–2349 (2021)
27. Jiang, Y., Wronski, B., Mildenhall, B., Barron, J., Wang, Z., Xue, T.: Fast and high-quality image denoising via malleable convolutions. Preprint. arXiv:2201.00392 (2022)

28. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711. Springer (2016)
29. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
30. Kim, J., Lee, S.: Deep learning of human visual sensitivity in image quality assessment framework. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1676–1684 (2017)
31. Kim, Y., Soh, J.W., Park, G.Y., Cho, N.I.: Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3482–3492 (2020)
32. Kong, F., Li, M., Liu, S., Liu, D., He, J., Bai, Y., Chen, F., Fu, L.: Residual local feature network for efficient super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 766–776 (2022)
33. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012)
34. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8878–8887 (2019)
35. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. *IEEE Access* **8**, 193907–193934 (2020)
36. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
37. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on Challenges in Representation Learning, ICML, vol. 3, p. 896 (2013)
38. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. *IEEE Trans. Image Process.* **28**(1), 492–505 (2018)
39. Li, J., Fang, F., Mei, K., Zhang, G.: Multi-scale residual network for image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 517–532 (2018)
40. Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., Yang, R.: Learning warped guidance for blind face restoration. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 272–289 (2018)
41. Li, L., Dong, Y., Ren, W., Pan, J., Gao, C., Sang, N., Yang, M.H.: Semi-supervised image dehazing. *IEEE Trans. Image Process.* **29**, 2766–2779 (2019)
42. Li, C., Guo, C., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. Preprint. arXiv:2103.00860 (2021)
43. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17452–17462 (2022)
44. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
45. Lin, X., Ma, L., Liu, W., Chang, S.F.: Context-gated convolution. In: European Conference on Computer Vision, pp. 701–718. Springer (2020)
46. Lin, Z., Garg, P., Banerjee, A., Magid, S.A., Sun, D., Zhang, Y., Van Gool, L., Wei, D., Pfister, H.: Revisiting rean: Improved training for image super-resolution. Preprint. arXiv:2201.11279 (2022)
47. Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9361–9370 (2021)

48. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. *IEEE Trans. Image Process.* **27**(6), 3064–3073 (2018)
49. Liu, J., Sun, Y., Xu, X., Kamilov, U.S.: Image restoration using total variation regularized deep image prior. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7715–7719. IEEE (2019)
50. Liu, J., Zhang, W., Tang, Y., Tang, J., Wu, G.: Residual feature aggregation network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2359–2368 (2020)
51. Lu, Z., Liu, H., Li, J., Zhang, L.: Efficient transformer for single image super-resolution. Preprint. arXiv:2108.11084 (2021)
52. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2502–2510 (2018)
53. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **21**(12), 4695–4708 (2012)
54. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal Process. Lett.* **20**(3), 209–212 (2012)
55. Nitzan, Y., Aberman, K., He, Q., Liba, O., Yarom, M., Gandelsman, Y., Mossner, I., Pritch, Y., Cohen-Or, D.: Mystyle: A personalized generative prior. Preprint. arXiv:2203.17272 (2022)
56. Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., Shen, H.: Single image super-resolution via a holistic attention network. In: European Conference on Computer Vision, pp. 191–207. Springer (2020)
57. Pan, J., Dong, J., Liu, Y., Zhang, J., Ren, J., Tang, J., Tai, Y.W., Yang, M.H.: Physics-based generative adversarial models for image restoration and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(7), 2449–2462 (2020)
58. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision, pp. 319–345. Springer (2020)
59. Ren, H., El-Khamy, M., Lee, J.: Dn-resnet: Efficient deep residual network for image denoising. In: Asian Conference on Computer Vision, pp. 215–230. Springer (2018)
60. Ren, X., Li, M., Cheng, W.H., Liu, J.: Joint enhancement and denoising method via sequential decomposition. In: 2018 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5. IEEE (2018)
61. Ren, W., Liu, S., Ma, L., Xu, Q., Xu, X., Cao, X., Du, J., Yang, M.H.: Low-light image enhancement via a deep hybrid network. *IEEE Trans. Image Process.* **28**(9), 4364–4375 (2019)
62. Richardson, W.H.: Bayesian-based iterative method of image restoration. *JoSA* **62**(1), 55–59 (1972)
63. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
64. Schuhmann, C.: LAION-Aesthetics. <https://laion.ai/blog/laion-aesthetics/>
65. Seif, G., Androutsos, D.: Edge-based loss function for single image super-resolution. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1468–1472. IEEE (2018)
66. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2107–2116 (2017)
67. Sungatullina, D., Zakharov, E., Ulyanov, D., Lempitsky, V.: Image manipulation with perceptual discriminators. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 579–595 (2018)
68. 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, pp. 1–9 (2015)

69. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3147–3155 (2017)
70. Talebi, H., Milanfar, P.: Nima: Neural image assessment. *IEEE Trans. Image Process.* **27**(8), 3998–4011 (2018)
71. Tian, C., Zhang, X., Lin, J.C.W., Zuo, W., Zhang, Y.: Generative adversarial networks for image super-resolution: A survey. Preprint. arXiv:2204.13620 (2022)
72. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. Preprint. arXiv:2204.01697 (2022)
73. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017)
74. Wang, Z., Bovik, A.C., Lu, L.: Why is image quality assessment so difficult? In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. IV–3313. IEEE (2002)
75. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
76. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
77. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6849–6857 (2019)
78. Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W.: Spatial attentive single-image deraining with a high quality real rain dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12270–12279 (2019)
79. Wang, Z., Zhang, J., Lin, M., Wang, J., Luo, P., Ren, J.: Learning a reinforced agent for flexible exposure bracketing selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1820–1828 (2020)
80. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9168–9178 (2021)
81. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1905–1914 (2021)
82. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. Preprint. arXiv:1808.04560 (2018)
83. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7268–7277 (2018)
84. Wei, W., Meng, D., Zhao, Q., Xu, Z., Wu, Y.: Semi-supervised transfer learning for image rain removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3877–3886 (2019)
85. Wei, P., Lu, H., Timofte, R., Lin, L., Zuo, W., Pan, Z., Li, B., Xi, T., Fan, Y., Zhang, G., et al.: Aim 2020 challenge on real image super-resolution: Methods and results. In: European Conference on Computer Vision, pp. 392–422. Springer (2020)
86. Wiener, N., Wiener, N., Mathematician, C., Wiener, N., Wiener, N., Mathématicien, C.: Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications, vol. 113. MIT Press, Cambridge, MA (1949)
87. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
88. Wu, G., Jiang, J., Liu, X., Ma, J.: A practical contrastive learning framework for single image super-resolution. Preprint. arXiv:2111.13924 (2021)

89. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10551–10560 (2021)
90. Xia, W., Zhang, Y., Yang, Y., Xue, J.H., Zhou, B., Yang, M.H.: Gan inversion: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* (2022)
91. Xu, D., Chu, Y., Sun, Q.: Moiré pattern removal via attentive fractal network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 472–473 (2020)
92. Xu, K., Yang, X., Yin, B., Lau, R.W.: Learning to restore low-light images via decomposition-and-enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2281–2290 (2020)
93. Xu, D., Poghosyan, H., Navasardyan, S., Jiang, Y., Shi, H., Wang, Z.: Recoro: Region-controllable robust light enhancement with user-specified imprecise masks. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 1376–1386 (2022)
94. Xu, X., Wang, R., Fu, C.W., Jia, J.: SNR-aware low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17714–17724 (2022)
95. Yang, W., Tan, R.T., Wang, S., Fang, Y., Liu, J.: Single image deraining: From model-based to data-driven and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(11), 4059–4077 (2020)
96. Yang, W., Wang, S., Xu, D., Wang, X., Liu, J.: Towards scale-free rain streak removal via self-supervised fractal band learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12629–12636 (2020)
97. Yang, T., Ren, P., Xie, X., Zhang, L.: GAN prior embedded network for blind face restoration in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 672–681 (2021)
98. Yoo, J., Ahn, N., Sohn, K.A.: Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8375–8384 (2020)
99. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. Preprint. arXiv:1511.07122 (2015)
100. Yu, K., Dong, C., Lin, L., Loy, C.C.: Crafting a toolchain for image restoration by deep reinforcement learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2443–2452 (2018)
101. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4471–4480 (2019)
102. Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 701–710 (2018)
103. Yuan, S., Timofte, R., Slabaugh, G., Leonardis, A.: Aim 2019 challenge on image demoiréing: Dataset and study. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 3526–3533. IEEE (2019)
104. Yuan, S., Timofte, R., Leonardis, A., Slabaugh, G.: Ntire 2020 challenge on image demoiréing: Methods and results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 460–461 (2020)
105. Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **20**(8), 2378–2386 (2011)
106. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
107. Zhang, X., Lu, Y., Liu, J., Dong, B.: Dynamically unfolding recurrent restorer: A moving endpoint control method for image restoration. Preprint. arXiv:1805.07709 (2018)

108. Zhang, X., Yang, W., Hu, Y., Liu, J.: Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 390–394. IEEE (2018)
109. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018)
110. Zhang, Z., Wang, X., Jung, C.: Dcsr: Dilated convolutions for single image super-resolution. *IEEE Trans. Image Process.* **28**(4), 1625–1635 (2018)
111. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning, pp. 7354–7363. PMLR (2019)
112. Zhang, Y., Wei, D., Qin, C., Wang, H., Pfister, H., Fu, Y.: Context reasoning attention network for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4278–4287 (2021)
113. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging* **3**(1), 47–57 (2016)
114. Zhao, H., Kong, X., He, J., Qiao, Y., Dong, C.: Efficient image super-resolution using pixel attention. In: European Conference on Computer Vision, pp. 56–72. Springer (2020)
115. Zhao, S., Zhang, L., Huang, S., Shen, Y., Zhao, S.: Dehazing evaluation: Real-world benchmark datasets, criteria, and baselines. *IEEE Trans. Image Process.* **29**, 6947–6962 (2020)
116. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

Chapter 16

Image Affect Modeling: An Industrial Perspective



Xin Lu

Abstract This chapter summarizes the formulation and industrial usages of image affect modeling in the past decade and aims to inspire thoughts to better scope and formulate the image affect modeling so that its value can be better realized in real-world applications. To bridge the gap between research and real-world applications for image affect modeling, I first present research questions that fall into this umbrella. Then industrial areas, such as content presentation, content creation, and camera and display, which image affect modeling is likely to be essential are discussed. Finally, I introduce two major opportunities for the use of image affect modeling, one related to multimodal analyses, and the other related to on-device learning involving rich contextual information. By broadly introducing these opportunities, I intend to motivate researchers and practitioners to join force to propose and solve problems in the image affect category.

16.1 Overview

Imaging intelligence has significantly transformed our world and lives in the past decade. Machines are intelligent enough to recognize objects and scenes, produce drawings and artwork, and drive cars and trucks. As a result, productivity has been significantly boosted in some of the labor-intensive industries (such as precision agriculture and port automated driving) and new forms of entertainment and creativity have been emerged, such as VR, vlogs (video logs), ChatGPT, AI-based text2image generation, text-based video editing, and AI-assisted social media apps. Among these breakthroughs, image recognition, segmentation, detection and generation are usually at the core of technology stacks and image affect modeling such as image aesthetics assessment and emotion classification are largely ignored.

X. Lu (✉)
Adobe Inc, San Jose, CA, USA
e-mail: xinl@adobe.com

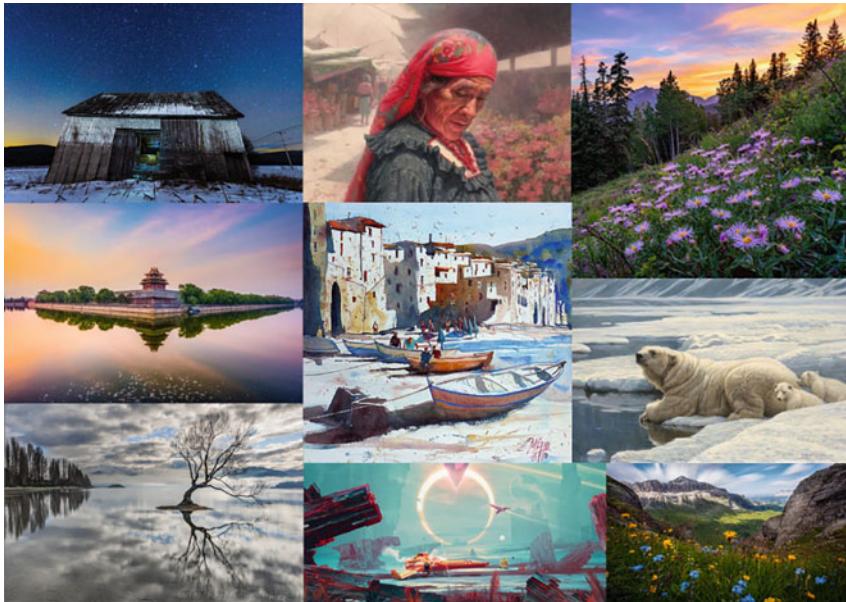


Fig. 16.1 Images from LAION-2B with the aesthetic score higher than 6.5 (Image ratings are from 1 to 10, and 10 is the highest rating) [14]

Is it useful to study the affect evoked in a human observer by an image (We refer to it as “image affect” in this chapter) after all from an industrial perspective?

In the past fifteen years, image affect modeling is mostly formulated as classification or regression problems [7, 16, 17]. The goal of a trained classifier or regressor is to predict how visually appealing in general the input image is or what types of emotions the image may evoke. Such tools have been widely adopted in many real-world applications, such as image search engine, personal album search and ranking, and image quality/aesthetics filtering. For example, when searching for images, you may notice that the returned image quality has been improved year over year. One reason that drives that improvement is the use of image aesthetics models [12, 21]. The effectiveness of the image aesthetics models is also demonstrated in recent text2image efforts such as stable diffusion [23]. According to the official announcements, the aesthetics predictor was used to filter training data and nearly half of the training iterations were done using the 600M highly aesthetic images (as shown in Fig. 16.1) in the LAION-2B dataset [15]. These real-world usages may not be noticeable to majority of the audiences, however, these success is huge to researchers and practitioners who have formulated, initiated, and contributed to image affect modeling. These success also indicate that image affect modeling provides additional information which is, to a certain extent, independent from image semantics. Identifying patterns from this aspect can compensate what has been learned so far from image semantics analyses.

Meanwhile, deep learning weighs heavily in the success of image affect modeling as it is known to be a mature tool to tackle classification and regression problems when the problem is well-defined and the volume of training data is sufficient. Therefore, image affect modeling is sometimes carelessly considered as a solved problem. Provided classification or regression were the optimal approach to formulate image affect modeling, image affect modeling can be considered as a solved problem in research communities. However, as researchers or practitioners in this field probably have noticed, even though images of high ratings and low ratings are generically agreed among us, some of them are debatable, and let alone the tremendous images that are rated in-between. This indicator suggests that the development of image affect modeling is far from mature. Whereas it is important to take the best technologies developed in image semantic analyses to image affect modeling, instead of continuing to improve the performance of image affect modeling as classification or regression problems, more efforts are needed to better scope this problem.

What are problems that image affect modeling is aiming for? I write this chapter to urge researchers and practitioners to think deeper and develop a unique set of problems under the umbrella of image affect modeling. For example, what are proper and effective computational models that can be used to learn from rich subjective responses of images or videos? Can it be learned to distinguish the circumstances that people are more likely to prefer A to B? Or can it be learned that people with certain characteristics are more likely to prefer A to B? These questions are largely ignored in image semantic analyses and instead may fit in the scope of image affect modeling well. Concrete scopes and problem definitions make this field an area where talented researchers and practitioners will be more willing to spend time and efforts. With increasing volumes of data and compute resources, the developed approaches are likely to impact and inspire other computational modeling tasks.

Image semantic analyses have mostly focused on answering the question of “what” and “where”, such as what this image/object is, where the dog is located in the image/video, and what makes Paris look like Paris [8]? Image affect modeling may focus on answering the question of “how” and “why”, such as how this dog is different from that dog in the image, how a sad dog differs from a happy dog, why a dog painted by a ten-year-old boy differs from the one made by a seventy-year-old gentleman, and why a boy may paint a happy dog one day and a sad dog the other day? In short, image affect modeling is far beyond rating images with scores, but it studies the association between visual characteristics and subjective opinions that are somewhat independent from image semantics under complex contexts and social-economic environments. Subjective opinions can be ratings, words or phrases, sentences or other forms of viewpoints.

To inspire thoughts, in the following of this chapter, a set of problems that I believe are valuable image affect modeling problems are presented. I first summarize some of my observations of image affect modeling and then introduce some of the industrial areas that are impacted by image affect modeling. Finally, two major opportunities for image affect modeling research and development are discussed.

16.2 Observations of Image Affect Modeling

Recent efforts in computational image affect modeling is commonly data-driven. Learning-based methods require data to construct mapping functions and optimization-based approaches use data to validate modeling or hypothesis. A known problem in image affect modeling is that data collection is far more challenging than other computer vision problems such as image recognition, detection, and segmentation. For example, in image recognition tasks, labelers are asked to label the content of an image. Assume the image is a dog, then majority of the labelers would consistently label that as a dog. However, if labelers are asked to rate how aesthetic or how happy the dog image is, then the distribution of scores usually varies largely (Fig. 16.2). The root cause is that the image affect modeling is contextual, subjective, and dynamic which can be modeled differently in different contexts, by different people, and at different times.

The image affect modeling is not a standalone problem, and it can manifest into various concrete problems in different contexts. For example, the aesthetic criteria for a nice oil painting is very different from a toddler's shot, which makes the comparison of the two challenging. Similarly, scenes captured with two cameras can both be aesthetic but of different looks. Therefore, rather than asking users to rate images by a score, can we model image affect on condition of the context? For example, instead of asking a question of “how do you rate the aesthetic value of this image”, can we ask “among sunset photos captured by Nikon within the proximity of this location, how do you rate the aesthetic value of this image”? The latter is more concrete and more likely to produce consistent labels. The plausibility of modeling complex context might be a concern in the past, and that's why researchers have attempted to analyze the impact of one factor at a time to estimate the contribution of a particular factor on image affect. That is a reasonable approach,



Fig. 16.2 Which image should be rated higher? Note: Both images are generated by AI. Sources of the images are <https://lexica.art/prompt/953a6608-9649-4aa9-ba69-5e379cf51088> and <https://lexica.art/prompt/06df188e-6767-4729-a225-80ef4e4ce509>

but not ideal. A breakthrough cannot be made without jointly modeling many factors because that's how the large volume real-world data is mostly generated. If we stick to data-driven approaches to uncover image affect rationales, we have to be courageous in handling complex contextual data.

Meanwhile, image affect modeling is subjective. Quantifying the idiosyncratic and shared contributions to judgment has been actively investigated in Psychology such as [18]. Those studies indicate that different people may have different criteria and metrics even if they are given the same context. For example, to take a portrait shot, some people prefer the look provided by the Google Pixel camera app and some people like the look provided by the iPhone camera app better. When these users label the image affect, conflicting labels occur. To model the subjective nature of this problem, one possibility is modeling the similarity of label distributions from different labelers and identifying typical distributions of labels. However, one labeler may have similar taste with another labeler on the lighting condition, but not on composition. In order to best leverage the training data, even within the same context, various attributes might need to be identified and modeled. If we model attributes jointly with subjectivity, given a portrait shot, a model can potentially conclude that labelers who value the lighting condition tend to rate this image high and labelers who value composition usually rate this image low. The real logic is likely to be more complex than this example, but modeling attributes jointly with subjectivity given contextual information is a plausible approach to figure out the nature of subjectivity.

Lastly, perception of image affect may evolve over time which has been demonstrated by the development of the fashion industry [19]. A visually appealing image in 1800s may not be appreciated as much today. To get more insights, these unique characteristics need to be incorporated in the problem formulation of image affect modeling. It would be interesting to figure out how our affective perception has evolved in the history.

16.3 Use Cases of Image Affect Modeling

Image affect modeling may have various use cases in different industries. In this section, the discussion is scoped to the software and the internet industry where image affect has been impacting nearly all the products and services.

UI/UX Design Software applications commonly require professional UI/UX design before shipping. The outputs of UI/UX design usually include icons, colors, layouts, fonts, transitions between pages and all other assets that are needed to present applications on the screen. Functionally, the design is expected to support product features, mission and vision. Aesthetically, the design is expected to delight users and emotionally connect with users. Design tools such as Sketch, Figma, and Adobe XD provide collaborative environments to quickly produce mockups of workflows and visual appeal of applications, but feedbacks of the mockups

are mostly collected manually through user studies. During the design process, although designers may get inspirations from images or mockups online and follow certain design patterns, most design decisions are based on the designers' subjective judgement. Designers wouldn't get immediate and diverse feedbacks from viewers' perspectives until the user-study phase. Therefore, the design process usually takes long and many rounds of iterations to incorporate comments and suggestions from developers, product managers, other designers and end users. When the application targets cross-platform and international deployment, the design process is even longer as users with different platforms or users of different cultures tend to have different visual preferences.

An image affect modeling tool in the UI/UX design space can potentially provide diverse insights instantaneously to designers and shorten the feedback loop. The inputs of the system include the product descriptions, initial designs, and rationales of the designs, and the outputs are predicted textual feedbacks from different perspectives. The system can easily be tailored to a specific domain or targeted audiences. In addition, an image affect modeling tool can also assist to verify the consistency across UI/UX designs of different pages in one application, designs across applications, and relevant marketing materials. Ultimately, a general AI might be able to draft UI/UX designs provided it were able to consume and analyze UI/UX designs.

Content Creation Content creation has a wider scope than UI/UX design. Whereas UI/UX design supports products functions, content creation usually aims to produce visually appealing content that conveys engaging, relevant, original and consistent messages. The UI/UX designs are mostly consumed by users of specific products, and the audiences of produced content ultimately can be everyone online. Similar with UI/UX designers, content creators conceptualize and visualize the messages through visuals and they need feedback from a more diverse group of audiences to finalize the creation. On the other hand, content creation involves images, videos, 3D models, vector graphics, text, audio and music. Such combination of multimedia requires an extended image affect modeling system that is able to understand multimodal inputs.

Content creation can either be photo-realistic, unrealistic, or a combination of both. For example, a poster may include artistic font together with some real pictures and emojis. Hence it requires a system that can handle inputs of different domains. Current deep learning approaches usually target analyzing one domain at a time. Researchers later adapt a model to different domains or train separate models for different domains. It remains an open question regarding whether to increase the network capacity to handle images of different domains or whether to chain a set of neural networks to handle cross-domain inputs. In general, analyzing content that involves various domains hasn't yet to be well formulated and tackled.

Meanwhile, content creation tool sets are diverse and the creation processes vary. Content creators start with blank pages, templates, existing images or visual components. And the outputs are highly unstructured. The same outputs can be produced by taking different steps in the same tool or using different tools. The



Fig. 16.3 AI Generated Images. Left: The prompt used is “a beautiful painting”. Right: The prompt used is “a very very very beautiful painting”

unstructured nature of content creation makes it a more challenging problem to provide step-by-step feedback compared to providing feedback on the completed artwork.

Recent breakthroughs of automatic image generation with text guidance, such as DALL-E [27], imagen [28], parti [31], midjourney [22], and stable diffusion [23], have aroused broad attention among designers, content creators, engineers, researchers, and business owners. For the first time, AI can consume human readable languages and visualize it with decent-quality images (with a resolution of 1024) at scale. Unlike images out of a search engine, images produced by these text2image tools mostly re-mix existing visual elements in the training dataset, rather than retrieve images from the training dataset. One observation of using these tools is that some of the affective and subjective words are modeled well, such as “very”. Images guided by the text prompt of “a very very very beautiful painting” tends to look aesthetically better than images guided by “a beautiful painting”, as shown in Fig. 16.3. Another observation is that the language modeling works better with nouns compared to adjectives. For example, images produced by “a sitting cat” tend to have a better visual quality than “a cat sitting”. More observations have even been included in the text prompt guideline produced by the community [6]. Considering the relationship between image affect and adjectives [9], which has been studied in the image affect modeling field, further investigation of the modeling quality towards image affect is interesting. Outcomes out of those studies would pave the way for text2affect, which is to generate affective and subjective expressions guided by languages. Recent development in language modeling [5, 25] and dialogue systems such as ChatGPT [24] and Bard [11] can augment affective expressions. As a natural extension, the proposed affective generation system can be applied to videos, audios, music, and other modalities.

Camera and Display Camera and display systems are on the other spectrum of content creation. The input of an image affect modeling system is usually 8-

bit tone-mapped images, i.e., the JPG images out of a camera, and the affective aspect of rendering a JPG image from RAW is often ignored. In addition to the sensor differences, different cameras have been equipped with different tone-mapping configurations. For example, Nikon is known to produce cooler colors while Canon is featured for a warmer look; the Pixel camera app tends to produce more natural look and the iPhone camera app produces a more HDR look. When comparing across different generations of iPhone cameras side by side, we are likely to notice differences as well. In the past, the tone-mapping configurations from OEM (Original Equipment Manufacturer), such as Apple, are mostly hidden from end users. Starting from iPhone 12 Pro (iOS 14.3 or later), users can shoot in the ProRAW format [2] where the tone-mapping configurations are partially revealed in the associated Profile Gain Table Map [1]. Such protocols open up the opportunity for researchers and practitioners to study the affective aspect of OEM's tone-mapping considerations. In the recently released iPhone 14 Pro (MAX), the wide camera (1x) produces 48MP ProRAW images (8064×6048 pixels) [2], which are four times larger than existing mobile photos (12MP, 4032×3024 pixels). The photo's resolution is 48MP only when shooting in ProRAW, not HEIC nor JPG, which provides additional incentive for people to shoot in ProRAW and therefore makes it possible for researchers to study the affective aspects of camera ISP from linear images.¹

Coupling with a camera's ISP system, another recently emerged area is EDR (Endpoint Detection and Response) display [29]. Apple talked about the EDR in the context of HDR (High Dynamic Range) implementation and HDR rendering in WWDC [4]. The reason that an EDR technique is emerging is because HDR means different things to different people in different scenarios. With the same content, it is likely to look differently on different hardware (iPad, iPhone, android phones, and laptops) and under different ambient conditions. For instance, the same content looks more vivid when viewed indoors than outdoors under natural sunlight. EDR thus aims to adapt the rendering algorithm to different situations and make the display best suited to a particular display and viewing environment. In a foreseeable future, developers might have more controls over EDR techniques, which means that opportunities for image affect modeling systems have been opened up in the camera and display industry.

Given the personal preferences and environmental conditions, an image affect modeling system may suggest the set of parameters that are customized and best suited. An image affect modeling system can also jointly consider the content creation, content presentation (UI/UX design) and the content display aspects to optimize for the content consumers.

¹ ProRAW is not a Bayer RAW image, which is a linear image in the RGB space.

16.4 Opportunities of the Use of Image Affect Modeling

Having presented the current status of image affect modeling and potential use cases in the industry, the biggest opportunities of the use of image affect modeling center at multimodal analyses and on-device learning.

Multimodal Analyses As image affect is contextual, subjective and dynamic, we need to incorporate those factors into the image affect modeling system to be able to identify patterns from data. Even though a picture is worth a thousand words, in many situations, an image itself is not sufficient to represent contexts, backgrounds, and preferences. For instance, an image may evoke positive emotional response because of a new born, and not because of its high aesthetics. An image may share negative messages when it captures a wild fire and again this doesn't mean the image itself is not aesthetic. To include these information into the modeling, inputs of other modalities such as text, usage logs, and audio need to be taken together with images to support learning.

The key idea of multimodal analyses is learning the association of features from different modalities on a granular level. In classification tasks, classes are usually represented through class identifiers in the algorithm. To generate affective comments or feedback, we need to train with pairs of images and affective comments. And the goal of the model training is learning and associating image features and text features. In the inference time, given an image, the model uses the learned association to generate text features and generate feedback in the form of text. Similarly, given text, the model uses the learned association to generate image features and generate images [22, 23, 27, 28, 31].

To incorporate contexts, instead of training with pairs of images and affective comments, we may train with pairs of images with textual descriptions of the content, context and affective comments. In the inference time, the model generates affective comments given an image with descriptions or generates images given textual descriptions. The textual descriptions of content and context can be replaced by other forms of information that convey the contexts.

On-Device Learning Contexts are essential for image affect modeling and contexts vary largely for different people in different tasks. A generic model might be able to tackle common situations or typical content creation problems, but it might always be necessary to adapt to each individual's situation. For an individual, local devices are centralized places that store these preferences. Customized preferences such as display brightness, preferred phone types and preferred apps, preferred document layouts, tags or notes associated with photos, usage logs and search queries are usually stored on local devices, such as laptop, phones, and tablets. Each individual application or cloud storage providers may have a piece of these data, but a complete view of customized preferences are on the local devices. Therefore, on-device learning is ideal to deliver a personalized image affect modeling system.

On-device learning doesn't necessarily mean to train a network from scratch on local devices that are not equipped with powerful GPUs. It could be in the form of

incremental learning [30], federated learning [13], or fine-tuning. Take the recently emerged stable diffusion model as an example. In [10], 3–5 additional images are used to adapt the released stable diffusion model to generate stylized images towards the given images. The personalization problem has not yet been formulated in the image affect modeling system, but on-device learning at least provides sufficient contextual data to learn from.

Meanwhile, hardware has been ramping up rapidly to support on-device training. Efforts to support accelerated training on dedicated devices have been making tremendous progress in the past five years. In particular, on Mac OS, Pytorch framework has announced its accelerated training using Metal Performance Shaders [26], i.e., Mac OS and iOS’s GPU compute shaders, which means deep neural network can train with GPUs on Mac OS, which is much faster than CPUs and SIMD accelerations. Apple’s ANE has supported model training since CoreML 3 [3]. It’s a matter of time to support typical deep neural network libraries like Pytorch with dedicated AI chip such as Apple’s Neural Engine. Additionally, mix-precision training configurations as well as training optimization tools such as deepspeed [20] make it possible to train deep neural networks with limited resources.

16.5 Conclusions

In this chapter, I discussed the achievement of image affect modeling in the past decades, the current progress, the biggest challenges and opportunities ahead. In particular, I introduced examples of industrial usages of the existing image affect models and explained three areas that have relied on or would heavily rely on image affect modeling systems, i.e., content presentation, content creation, and camera and display. I dived into the fundamentals of image affect modeling problems by comparing to image classification, detection, and segmentation problems. By providing a broad overview as such, I am hoping to inspire readers to think creatively on the problem scopes and formulations of image affect modeling and contribute novel solutions.

References

1. Adobe Inc.: Digital negative (dng) specification, version 1.6.0.0. https://helpx.adobe.com/content/dam/help/en/photoshop/pdf/dng_spec_1_6_0_0.pdf. Online. Accessed 23 Sept 2022
2. Apple Inc.: About Apple ProRAW. <https://support.apple.com/en-us/HT211965>. Online. Accessed 23 Sept 2022
3. Apple Inc.: Core ML 3 Framework. <https://developer.apple.com/videos/play/wwdc2019/704/>. Online. Accessed 23 Sept 2022
4. Apple Inc.: Explore HDR rendering with EDR. <https://developer.apple.com/videos/play/wwdc2021/10161/>. Online. Accessed 23 Sept 2022

5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
6. dallery.gallery: Dalle 2 prompt book. <https://dallery.gallery/wp-content/uploads/2022/07/The-DALL%C2%B7E-2-prompt-book-v1.02.pdf>. Online. Accessed 23 Sept 2022
7. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 288–301 (2006)
8. Doersch, C., Singh, S., Gupta, A., Sivic, J., Efros, A.A.: What makes Paris look like Paris? ACM Trans. Graph. (SIGGRAPH) **31**(4), 101:1–101:9 (2012)
9. Fernandez, D., Woodward, A., Campos, V., Giro-i Nieto, X., Jou, B., Chang, S.F.: More cat than cute? Interpretable prediction of adjective noun pairs. In: Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes (MUSA2), Proceedings of the ACM Multimedia Conference (2017)
10. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv (2022). <https://doi.org/10.48550/ARXIV.2208.01618>. <https://arxiv.org/abs/2208.01618>
11. Google: Bard. <https://blog.google/technology/ai/bard-google-ai-search-updates/>. Online. Accessed 15 Feb 2023
12. Google AI Blog: Introducing NIMA: Neural image assessment. <https://ai.googleblog.com/2017/12/introducing-nima-neural-image-assessment.html> (2017). Online. Accessed 23 Sept 2022
13. Konečný, J., McMahan, H.B., Yu, F.X., Richtarik, P., Suresh, A.T., Bacon, D.: Federated learning: Strategies for improving communication efficiency. In: NIPS Workshop on Private Multi-Party Machine Learning (2016). <https://arxiv.org/abs/1610.05492>
14. LAION-AI: Laion aesthetics. <https://laion.ai/blog/laion-aesthetics/> (2022). Online. Accessed 23 Sept 2022
15. LAION-AI: Laion datasets. <https://github.com/LAION-AI/laion-datasets> (2022). Online. Accessed 23 Sept 2022
16. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.: Rapid: Rating pictorial aesthetics using deep learning. In: Proceedings of the ACM Multimedia Conference, pp. 457–466 (2014)
17. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.: Rating pictorial aesthetics using deep learning. IEEE Trans. Multimedia **17**(11), 2021–2034 (2015)
18. Martinez, E.J., Funk, F., Todorov, A.: Quantifying idiosyncratic and shared contributions to judgment. Behav. Res. Methods **52**, 1428–1444 (2020)
19. McRoberts, L. B.: Petite women: Fit and body shape analysis. thesis of master of science (2005). Thesis of Master of Science, The School of Human Ecology, Louisiana State University. Online. Accessed 23 Sept 2022
20. Microsoft: Deepspeed. <https://github.com/microsoft/DeepSpeed>. Online. Accessed 23 Sept 2022
21. Microsoft Bing Blog: Enhancing image quality in microsoft bing. <https://blogs.bing.com/search-quality-insights/september-2021/enhancing-image-quality-in-microsoft-bing>. Online. Accessed 23 Sept 2022
22. Midjourney: <https://www.midjourney.com/home/>. Online. Accessed 23 Sept 2022
23. Mostaque, E.: Stable diffusion public release. <https://stability.ai/blog/stable-diffusion-public-release>. Online. Accessed 23 Sept 2022
24. OpenAI: ChatGPT. <https://openai.com/blog/chatgpt/>. Online. Accessed 15 Feb 2023
25. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. arXiv (2022)

26. PyTorch: Introducing accelerated PyTorch training on Mac. <https://pytorch.org/blog/introducing-accelerated-pytorch-training-on-mac/>. Online. Accessed 23 Sept 2022
27. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. CoRR abs/2102.12092 (2021). <https://arxiv.org/abs/2102.12092>
28. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. Arxiv (2022). <https://doi.org/10.48550/ARXIV.2205.11487>. <https://arxiv.org/abs/2205.11487>
29. Singh, D.: What is Apple EDR? How is it different from regular HDR? <https://www.digit.in/features/general/apple-edr-how-is-it-different-from-regular-hdr-59940.html>. Online. Accessed 23 Sept 2022
30. Utgoff, P.E.: Improved training via incremental learning. In: Proceedings of the Sixth International Workshop on Machine Learning, pp. 362–365 (1989)
31. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation. Arxiv (2022). <https://arxiv.org/abs/2206.10789>

Part VI

Emotion

This part of the book specifically examines how felt emotion influences aesthetics, and on the expression and perception of emotion in the face and body.

The first chapter, “Emotional Expression as a Means of Communicating Virtual Human Personalities,” examines how “virtual” humans can convey personality by expressing different emotions through the face and body. This chapter highlights the importance of emotion communication in our interactions with virtual humans.

The second chapter, “Modeling Emotion Perception from Body Movements for Human-Machine Interactions using Laban Movement Analysis,” then extends our understanding of computer vision of facial expressions of emotions to bodies. It presents work applying the Laban Movement Analysis to train computer models to read specific emotions from bodily movements.

The third chapter, “Demographic Differences and Biases in Affect Evoked by Visual Features,” then examines how one’s felt emotion in response to pictorial scenes influences aesthetic judgments and focuses on perceiver characteristics such as gender and culture.

Chapter 17

Emotional Expression as a Means of Communicating Virtual Human Personalities



Sinan Sonlu, Khasmamat Shabanovi, Uğur Güdükbay, and Funda Durupinar

Abstract Virtual humans with realistic behaviors have become prominent actors of compelling virtual experiences in domains as diverse as entertainment, education, and healthcare. A significant factor contributing to their behavioral realism is their personality, which characterizes distinctive traits consistent over time. Virtual humans can express personality traits through various channels such as voice, face, or body. In this chapter, we will focus on how emotional expression through facial expressions and body pose affect the communication of virtual human personalities. Throughout the chapter, we refer to the five-factor model of personality, which consists of five orthogonal dimensions of openness, conscientiousness, extroversion, agreeableness, and neuroticism. We will investigate their representation through the expression of the basic emotions of happiness, sadness, fear, anger, and disgust.

17.1 Introduction

Whether for an ant building the colony's nest or a lion hunting for food, communication is essential to an organism's survival. As much more complicated social beings, humans communicate with the purpose of more than just a message transfer between sides. We exchange feelings and desires, expressing our inner selves through intricate verbal and non-verbal signals. Despite the signal complexity and independent of the closeness of the relationship, understanding the feelings

S. Sonlu · U. Güdükbay

Department of Computer Engineering, Bilkent University, Ankara, Turkey

e-mail: sinan.sonlu@bilkent.edu.tr; gudukbay@cs.bilkent.edu.tr

K. Shabanovi

Technical University of Munich, Munich, Germany

F. Durupinar (✉)

Department of Computer Science, College of Science and Mathematics, University of Massachusetts, Boston, MA, USA

e-mail: funda.durupinarbabur@umb.edu

of another individual is straightforward for an average person. We unconsciously analyze the facial expressions, gestures, and vocal signals of the person we communicate with and make instant deductions about their emotions. Emotional signals have universal associations and can be recognized similarly across cultures or even species [20, 27]. This knowledge allows actors to replicate signals of affect, regardless of whether they have internalized the corresponding feelings or not [81]. Such knowledge also informs the design of virtual characters: although virtual characters do not “feel”, their accurate manifestation of emotional signals makes them believable and engaging.

Virtual characters are essential to the digital world, from video games and animated films to virtual assistants and social avatars. We expect these characters to look and act like humans, exhibiting consistent behaviors representative of their characteristics. Consistency and human-like behavior can be established by imbuing personality into virtual characters as personality quintessentially defines an individual’s long-term and distinctive traits. Additionally, multiple studies have shown that people can accurately assess the personalities of virtual humans based on verbal and non-verbal cues [25, 42, 45, 70, 85]. In this chapter, we examine how virtual humans’ emotional facial expressions and body postures affect the perception of their personalities.

Ample research indicates that personality traits control the intensity and frequency of emotional responses [55, 61, 63, 69]. Even without a causal link, we associate certain emotional responses with specific personality traits [13, 85]. First impressions of personality are often influenced by the perceived emotional content of an individual’s neutral facial expression, which is controlled by physical features such as the shape of the face [1]. This influence carries the risk of overgeneralization and stereotyping, as certain features may be attributed to specific emotions and genders. For example, rounder faces are often associated with females and resemble fear and surprise expressions, leading to the perception of submissiveness. Without falling into the trap of stereotyping, this chapter presents our research on the relationship between the usage of facial expressions and body poses, rather than physical features, that indicate emotions and their impact on personality perception.

For personality description, we use the five-factor model [23], which is an established and widely-adopted personality model. The five factors are openness, conscientiousness, extroversion, agreeableness, and neuroticism. For emotions, we refer to the six basic emotions of sadness, happiness, anger, fear, surprise, and disgust [46]. Although there are contradicting theories on which emotions are universal [28, 80], these six are commonly recognized in the Western world.

We describe two user studies to collect judgments on virtual humans’ apparent personalities, given images depicting them with emotional facial expressions and body poses. The results indicate statistically significant links between the participants’ perception of virtual human personalities and their emotional expressions. For example, we found sadness related to introversion and anger to low agreeableness. We found that facial expression and body pose determine different personality factors. For instance, agreeableness was better represented in facial expressions, whereas emotional poses were more indicative of extroversion.

Body poses that express the same emotion have higher variance than facial expressions, which directly correspond to emotions; consequently, we take a closer look at the subtle pose differences that cause personality shifts. We also discuss how pose descriptors based on Laban Movement Analysis (LMA) can explain these personality shifts.

Controlling the emotional content of a virtual character in line with these findings can help animators and researchers design better personality expressions and more realistic communication. We should note that the interpretation of emotional expression may vary without contextual information [64]. In such cases, different signals such as body pose [9], lighting conditions [100], or the background scene [98] can provide context and improve the accuracy of emotional perception. However, in this chapter, we analyze emotions in isolation from context, for the sake of facilitating controlled experiments by reducing the number of variables involved.

17.2 Background

The five-factor model of personality is supported by a considerable number of empirical studies [89] that explain its cross-cultural coverage [48], neurobiological correlates [22], temporal stability across the life span [78], and genetic structure [99]. The model investigates the psychological nature of an individual under five orthogonal dimensions [23], each grouping multiple traits on a bipolar and continuous scale. Openness measures curiosity and creativity. People with high openness tend to enjoy trying new experiences; in contrast, individuals with low openness dislike change. Conscientiousness is related to controlling and planning. High conscientiousness relates to being organized; people with low conscientiousness tend to act irresponsibly. Extroversion examines the social aspect of interaction. Extroverted individuals tend to be more outgoing and energetic, while introversion is associated with being reserved. Agreeableness measures empathy. High agreeableness relates to being understanding and kind; low agreeableness involves rude and irresponsible behavior. Neuroticism examines the tendency towards anxiety and negative feelings. High neuroticism is associated with anxious behavior, while low neuroticism corresponds to being calm and secure.

The face is the focal point of interpersonal communication. We are evolutionarily attuned to facial expressions [33] as they have evolved from our needs. For instance, we express pain to request sympathetic attention [96] and fear to alert others [91]. A combination of facial muscles acts together to form an expression. Rather than focusing on individual muscle activities, Facial Action Coding System (FACS) [29] describes facial expressions where each atomic movement on the face is classified with an Action Unit (AU). For example, the facial expression of joy includes AU 6 (Cheek Raiser) and AU 12 (Lip Corner Puller). The omission of an AU influences the genuineness of the corresponding expression—a happy expression without AU 6 is more likely to be perceived as fake. The realism of the facial expressions of the computer-generated characters highly depends on the correct usage of AUs.

Primary emotions are strongly related to facial expressions with well-defined parameter combinations. Body poses also convey emotions, although their link is not defined precisely. For example, rising and spreading joints can signal happiness, but there is no universally accepted single posture to show happiness, unlike the facial parameters of happiness. Introducing context, such as placing a gift box in front of a person to convey happiness, can improve the understanding of emotions. However, in this chapter, our focus is on analyzing the influence of different body poses on personality perception, without the presence of contextual factors. Thus, evaluating a set of poses regarding a specific emotion is more meaningful than using a single pose.

A possible choice to associate individual joints' contribution to the emotional content of a pose is to utilize LMA. LMA [38] offers a formal system to describe, visualize and interpret human movement under four main categories: Body, Effort, Shape, and Space. The Body category describes the structural and physical attributes of the human body during movement. The body parts involved in the movement and their influence on other body parts are examined under this category. Effort defines the dynamic characteristics of movement concerning inner intention. The difference between an angry punch and a friendly tap is identified by four Effort Factors (Space, Weight, Time, and Flow). Shape expresses the way the body changes shape during movement. Shape Qualities, a subcategory of Shape, describe this change relative to a spatial reference point (Rising/Sinking, Spreading/Enclosing, and Advancing/Retreating). Finally, Space examines the motion in connection with the environment.

17.3 Related Work

Research on expressive virtual humans spans various fields with the ultimate goal of creating human-like behavior. Accurate representation of emotions and personality is a part of achieving this goal. Under these two categories of affect, we discuss research focusing on recognition and expression/synthesis. Although our main objective is to establish expressive communication, recognition is essential for uncovering the factors contributing to successful expression.

17.3.1 *Emotion and Personality Recognition*

Works that study emotion recognition generally focus on the face and body [3, 10, 16, 40, 47]. Wegrzyn et al. [95] investigate the influence of each facial region on emotion recognition by revealing one sub-region until the participant decides that the expression is recognizable. They report that sadness and fear are manifested in the eyes while disgust and happiness in the mouth. They devise and validate

a correlation between facial action units and the expressions that they describe. Glowinski et al. [35] analyze the upper body motion to relate actor movements with four emotional categories of anger, joy, relief, and sadness, utilizing geometric descriptors based on the triangle formed by the head and the hands. They employ features inspired by LMA, such as calculating the convex hull volume of the body joints to describe the Shape component [4]. Parameters that describe movement style can be used for identifying emotions [15, 18, 39, 76].

Although face and body play the most critical roles in emotion recognition, speech [56], psychological signals [82], and text [6] also give extensive information. Context can improve the accuracy of emotion recognition by providing additional information to the otherwise ambiguous emotions [9, 98, 100]. The most successful approaches utilize multiple audiovisual features in deep architectures [93].

Apparent personality can be estimated based on motion cues [57], facial videos [26], body shape [43], social network profiles and messages [31, 50], physiological sensor data [86], and portrait images [71]. Even an individual's room can reveal cues about their personality [36]. Multi-modal systems that combine numerous cues yield superior performance in personality recognition [8, 11]. Using LMA-based features improves emotion [7] and personality [30] recognition from the skeletal pose. The success of affect recognition, which relies on data-driven methods, is often dependent on the availability of a large data set. Therefore, many studies use in-the-wild data sets in order to increase the sample size [64]. Such data sets introduce higher variance compared to acted data sets acquired in controlled environments. Because we are interested in discovering the personality connections of different emotional poses, we require each sample to have strong emotional associations. To this end, we utilize the BEAST data set [21] that includes acted poses with clear emotional associations.

17.3.2 *Emotion and Personality Synthesis*

Existing work mostly focuses on the facial expression of emotions [41, 44, 54, 74, 97] as the lack of facial expressions can cause an uncanny effect, dramatically reducing a virtual human's plausibility [90]. For a facial expression to be recognizable, the virtual character should include sufficient detail signaling the emotion [12]. When the faces of the individuals are not visible [62], such as crowd simulations [24], certain patterns of body motion help portray specific emotions [19]. When the emotions conveyed by the face and body cooperate, they improve communication [87]. However, when body and face express conflicting emotions, body language is more influential than the facial expressions on observers' judgments [68].

LMA is used to control high-level motion parameters that govern the style of human motion. Through automatic adjustments that affect these parameters, generative systems can convey different emotions using the same input motion [14]. To this end, qualitative LMA elements are translated into quantitative motion features

utilizing empirical frameworks [17]. Designing the character’s motion using such LMA-based quantitative attributes is highly beneficial in gesture animation [5, 25]. These attributes are also used in expressing emotions in robot motion where physical constraints are more restrictive [67].

Gestures can help express different personalities in human-like robots and virtual characters [53, 73, 84]. Motion of the hands [94], the use of facial expressions [85], voice style [75], and dialogue content [66] are all important factors that influence personality expression.

In this work, we focus on static emotional facial expressions and poses and leave the analysis of animations as future work. The perceived intensity of emotional expressions for static images is slightly less than their dynamic counterparts [52], but emotion recognition accuracy in static and dynamic images are mostly similar [51]. Consequently, we investigate the interactions in the static space where fewer variables are involved. We expect a similar but stronger response in dynamic emotional facial expressions and poses. Nevertheless, static emotional faces and poses are heavily used in websites, illustrated books, and visual novels due to their lower cost and the limitations of the media.

Recent work in synthesizing novel poses using human images [34, 60, 83] is promising for generating realistic virtual agent imagery. Combined with methods of facial expression transformation [65, 88], a single image can span many frames expressing a wide range of emotions. Generative Adversarial Network (GAN)-based architectures can be an alternative to the tiresome process of creating realistic 3D humans. Generative networks can also produce novel poses that exhibit the target emotions [2]. We follow such an approach for body pose generation.

17.4 The Effect of Emotions on Personality Perception

Devising a one-to-one mapping between emotions and personality is impractical, as people can feel and express the same emotions regardless of their personalities. On the other hand, a large body of research acknowledges some personalities’ increased susceptibility to specific emotions and their control of emotional expressivity [49, 77]. In this chapter, we explore how this knowledge applies to virtual humans and to what extent emotions expressed through the face and body pose impact the perception of the five personality factors.

We describe two user studies designed to find associations of specific image categories with the apparent personality factors they suggest. The first study investigates the effect of emotional facial expressions of 3D human models, and the second study analyzes the impact of emotional poses of synthetic images on apparent personality.

17.4.1 Study 1: Emotional Facial Expressions and Personality Perception

For the first study, we designed the facial expressions of happiness, sadness, anger, fear, and disgust on a 3D model using Adobe Fuse. We tuned the facial blend shapes of the model according to the FACS specification of AU intensities. We captured six images of the model with neutral, happy, sad, angry, scared, and disgusted expressions (cf. Fig. 17.1). We did not include surprise as it was indistinguishable from the scared expression, given the facial blend shapes of our model. Because of the universal recognition and precision of the facial expressions, we prepared one image per category.

We conducted an Amazon Mechanical Turk study to collect judgments on the perceived personality factors in each image. The virtual human's physical appearance was the same; only the facial expressions changed across each image. We asked participants to rate the personality of the character on the image 7-point Likert scale [58] using Ten Item Personality Inventory (TIPI) [37], which is a validated, brief personality inventory. At the beginning of the study, we showed a set of facial expressions to prepare the participant. The participant was allowed to view each sample without any time limitation. Samples were shown on the screen one at a time in random order. Each image was evaluated by 100 individuals (64 male and 36 female, with a mean age of 29.4).

We grouped the results based on their emotion labels to analyze the relationship between the emotion category and the perceived personality. Following the standard usage of TIPI, we averaged the participants' choices to the related questions to calculate the per-factor personality score. The Likert-scale OCEAN score distributions of each emotional category per personality factor are shown in Fig. 17.2.



Fig. 17.1 Samples with different facial expressions used in the first study

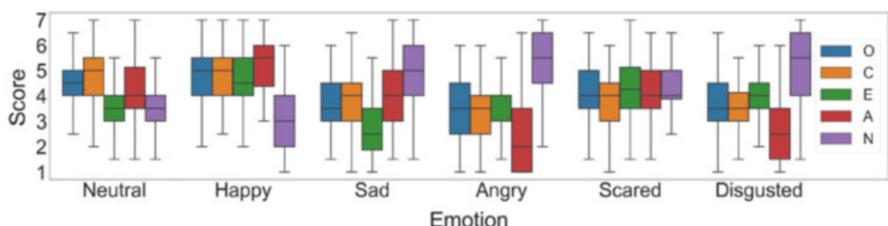


Fig. 17.2 Likert scale OCEAN score distribution of each facial expression in the first study

We observe that the neutral expression portrays subtle traits and does not correspond to a neutral personality (Fig. 17.2). The lack of a facial expression may have led the participants to focus on the character's physical appearance, which can also convey signals about personality [43]. Various traits can influence the perception of a general personality dimension, called the Big One [72, 79]. This is a phenomenon where all personality factors collapse into a single dimension, and are observed as either positive or negative (neuroticism being reversed, as high neuroticism has negative connotations). We observe a general positiveness in happy facial expressions. For the negative pole, we find that high neuroticism is common in sad, angry, and disgusted emotions, while introversion is linked to sadness, and disagreeableness is related to anger and disgust.

The happy expression scores highest in openness, conscientiousness, extroversion, and agreeableness. The angry expression has the lowest openness, conscientiousness, and agreeableness scores, whereas the sad expression has the lowest extroversion score. The highest neuroticism score is achieved by anger, and the lowest score by happiness. Images or animations of virtual humans portraying these emotional facial expressions can communicate the corresponding personalities. Because the personality scores of neutral and scared facial expressions are weak, they can depict neutral traits.

For the emotions highly influential on the perception of a specific factor, we expect more divergence from neutral personality, which corresponds to a score of 4 on a 7-point Likert scale. For example, angry and sad expressions both indicate neuroticism; however, the mean neuroticism score of anger is closer to 7, making it a better candidate for expressing neuroticism.

The highest variance is for agreeableness and neuroticism, while the lowest is for openness and conscientiousness. Low variance in openness and conscientiousness is also observed in similar research [85], possibly because these traits are hard to observe in a short time.

For further analysis, we calculated each image's average scores per personality factor and grouped them based on their emotion category. Then, we performed an ANalysis Of VAriance (ANOVA) per personality factor to evaluate the statistical significance of the differences between the mean scores of emotion categories. The null hypothesis assumes no statistically significant difference. Since ANOVA only reports the existence of statistical significance, we also performed Tukey's HSD [92] to find significantly different means.

The mean differences on a 7-point scale are shown in Table 17.1. The colored cells highlight the statistically significant differences. For each emotion pair and factor of interest, the mean difference was calculated by subtracting the mean score of the emotion on the right-hand side from the one on the left. If the mean score difference is positive, the emotion on the left-hand side has a higher score than the emotion on the right.

The highest mean differences are achieved for agreeableness and neuroticism between happy-angry and happy-disgusted pairs. Compared to the neutral expression, the highest difference is achieved by the angry expression, while the lowest is for the happy expression. The happy and sad emotions have opposite signs for each

Table 17.1 Facial expression ANOVA study results. For each emotion pair, the mean difference was found by subtracting the mean score of the emotion on the right-hand side from the mean score of the emotion on the left, where per factor mean scores were calculated in terms of the 7-point Likert scale. Colored cells show statistically significant differences. Gray shows differences up to 1; blue shows differences between 1 and 2, and yellow shows differences higher than 2

Emotion pair	Δ_O	ρ_O	Δ_C	ρ_C	Δ_E	ρ_E	Δ_A	ρ_A	Δ_N	ρ_N
Neutral – happy	-0.345	0.243	-0.205	0.765	-0.945	0.001	-0.935	0.001	0.445	0.094
Neutral – sad	0.725	0.001	0.770	0.001	0.785	0.001	0.200	0.843	-1.420	0.001
Neutral – angry	1.06	0.001	1.22	0.001	-0.220	0.741	1.870	0.001	-1.985	0.001
Neutral – scared	0.190	0.812	0.880	0.001	-0.820	0.001	-0.020	0.900	-0.885	0.001
Neutral – disgusted	0.765	0.001	1.050	0.001	-0.375	0.209	1.675	0.001	-1.740	0.001
Happy – sad	1.070	0.001	0.975	0.001	1.730	0.001	1.135	0.001	-1.865	0.001
Happy – angry	1.405	0.001	1.425	0.001	0.725	0.001	2.805	0.001	-2.430	0.001
Happy – scared	0.535	0.009	1.085	0.001	0.125	0.900	0.915	0.001	-1.330	0.001
Happy – disgusted	1.11	0.001	1.255	0.001	0.570	0.008	2.610	0.001	-2.185	0.001
Sad – angry	0.335	0.275	0.450	0.055	-1.005	0.001	1.670	0.001	-0.565	0.012
Sad – scared	-0.535	0.009	0.110	0.900	-1.605	0.001	-0.220	0.777	0.535	0.021
Sad – disgusted	0.040	0.900	0.280	0.494	-1.160	0.001	1.475	0.001	-0.320	0.415
Angry – scared	-0.870	0.001	-0.340	0.271	-0.600	0.004	-1.890	0.001	1.100	0.001
Angry – disgusted	-0.295	0.422	-0.170	0.892	-0.155	0.900	-0.195	0.860	0.245	0.676
Scared – disgusted	0.575	0.004	0.170	0.892	0.445	0.079	1.695	0.001	-0.855	0.001

factor. Anger and disgust are perceived very similarly, but negative associations with disgust are slightly less than with anger. The highest mean difference total is between happy and angry expressions. Sadness differs from other emotions of negative connotation (anger, fear, and disgust) in terms of extroversion, as sadness is more associated with introversion while others are with extroversion. This aspect of sadness can help isolate extroversion to control the apparent personality better.

17.4.2 Study 2: Emotional Body Poses and Personality Perception

The second study investigates the relationship between emotional poses and apparent personality. In contrast to facial expressions, the emotional content of different body poses is not universally recognized. Therefore, we represented each category with multiple images in this study. We generated 40 full-body images from four emotional categories (angry, happy, sad, and scared) using Liquid Warping GAN [60], a pre-trained pose transfer network. Liquid Warping GAN is a multi-task model that can be used for *human motion imitation*, *novel view synthesis*, and *appearance transfer*. We used human motion imitation that takes a source image and a target pose to generate a novel image of the source expressing the target pose. Each emotional category included ten pose variants to compensate for the variance in poses representing emotions.

We generated full-body images from the emotional poses in BEAST database [21], which includes 254 poses produced by 46 individuals expressing four emotions. Images in the BEAST database lack the faces of the actors, which occasionally causes pose estimation failure in specific images. To overcome this issue, we imitated a subset of 40 poses in the database. For the input source images that represent the body appearance, we utilized the DeepFashion [59] data set. Figure 17.3 shows three images generated from each emotion category. Unlike the first study, the categories of neutral and disgusted were not included since they were not available in the reference database, possibly because they lacked clear representations of body poses.

We conducted an online user study where 35 participants (25 male and 10 female, mean age of 21.6) rated the apparent personality in each image using TIPI [37]. Due to the increased sample size in this study, we used a 5-point Likert scale [58].

Like the first study, we showed a set of sample poses for warm-up at the beginning. The participant was allowed to view each sample without any time limitation. Samples were shown on the screen one at a time in random order. Participants of the two studies were non-overlapping.

We grouped the results based on their emotion labels to analyze the relationship between the emotion category and the perceived personality. The Likert-scale OCEAN score distributions of each emotional category per personality factor are shown in Fig. 17.4.

The results of the pose study indicate less divergence from the neutral personality (cf. Fig. 17.4). This divergence could be caused by the multiple images representing each emotional category. Subtle changes in the pose could result in different personalities, and grouping such poses together could have diminishing effects. In this case, a closer look at each pose can reveal exciting results, which we leave



Fig. 17.3 Pose samples from each emotional category used in the second study

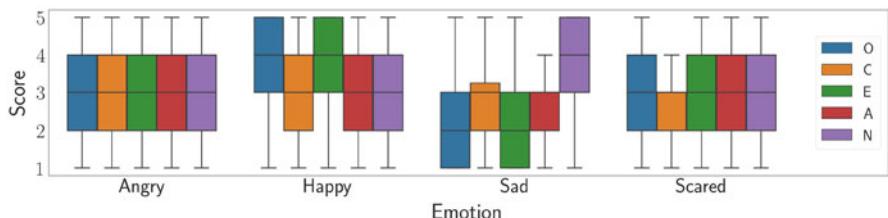


Fig. 17.4 Likert scale OCEAN score distribution of each emotional category of the pose study

Table 17.2 Pose transfer ANOVA study results. For each emotion pair, the mean difference was found by subtracting the mean score of the emotion on the right-hand side from the mean score of the one on the left. Per-factor mean scores were calculated in terms of the 5-point Likert scale. Colored cells indicate statistically significant differences. Gray indicates differences up to 1, and blue indicates differences higher than 1

Emotion pair	Δ_O	ρ_O	Δ_C	ρ_C	Δ_E	ρ_E	Δ_A	ρ_A	Δ_N	ρ_N
Happy – angry	0.301	0.459	-0.071	0.900	0.547	0.187	0.322	0.151	-0.094	0.900
Sad – angry	-0.828	0.001	-0.292	0.130	-1.198	0.001	-0.162	0.676	0.626	0.007
Scared – angry	-0.147	0.880	-0.314	0.092	-0.122	0.900	-0.053	0.900	0.205	0.649
Sad – happy	-1.129	0.001	-0.221	0.338	-1.745	0.001	-0.489	0.012	0.720	0.002
Scared – happy	-0.448	0.141	-0.243	0.258	-0.669	0.075	-0.376	0.072	0.299	0.355
Scared – sad	0.681	0.010	-0.022	0.900	1.076	0.002	0.109	0.875	-0.421	0.107

for future studies. When we group the samples based on emotion, we observe that angry and scared poses are perceived as more neutral. We find that happy poses indicate high openness and extroversion. Sad poses, on the other hand, convey the opposite traits in addition to high neuroticism. Similar to the first study, a tendency to perceive the general positiveness [72] is prominent.

The results are shown in Table 17.2 on a 5-point scale. The colored cells highlight the statistically significant differences. The mean differences were calculated similar to the study of facial expressions.

The highest mean differences are achieved for extroversion, while different emotional poses do not significantly differ in conscientiousness. We observe the highest mean difference between sad and happy poses, while the most similar personalities are found in scared and angry poses. Similar to the first study, sadness is highly related to introversion. Sad poses are also low in openness compared to the other emotions.

We also compare the personality scores of various individual poses because subtle variations in pose can result in significant personality differences. We only look at a few examples of pose couples due to limited space, but the complete results of our user study are available at <https://github.com/khasmamad99/personalityTransfer> for further research.

In Fig. 17.5, we compare two angry poses. The figure on the left has his hands on his hips. In contrast, the figure on the right keeps his hands together with a slightly turned posture and tilted head. When the figure's body is not directed toward the camera, we observe an increase in agreeableness.

In Fig. 17.6, we compare two happy poses. The figures primarily differ in terms of hand positions. The figure on the left has a spreading pose, while the figure on the right keeps his hands closer to his body. The figure on the right has a slightly wider foot positioning. We observe an increase in the general positiveness dimension when the pose spreads more. The most influenced factors are extroversion and openness.

In Fig. 17.7, we compare two sad poses. While the two figures are mostly similar, the figure on the left is looking down. In contrast, the figure on the right looks directly at the camera and is slightly more rising. We observe a significant

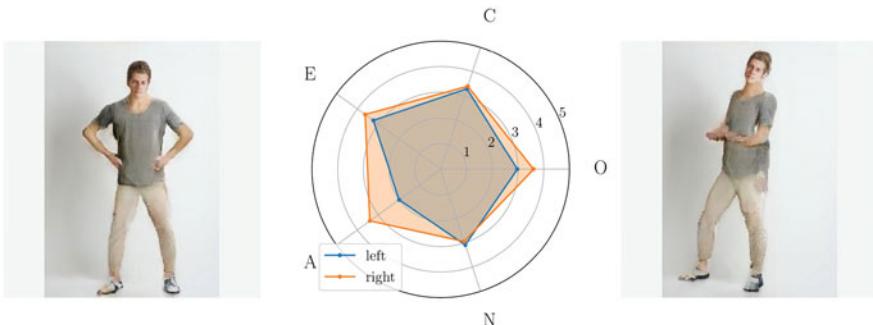


Fig. 17.5 Comparison of two angry poses

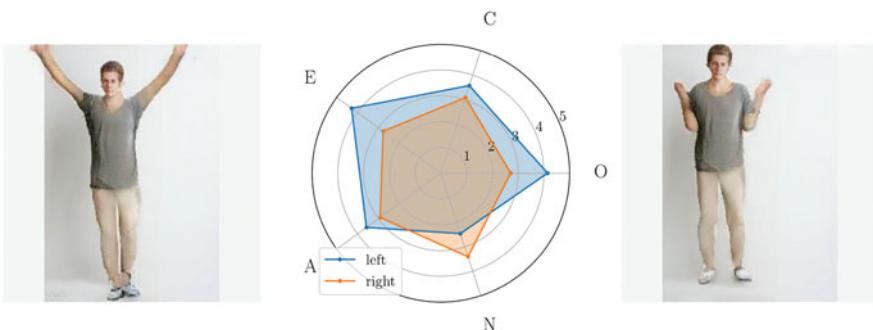


Fig. 17.6 Comparison of two happy poses

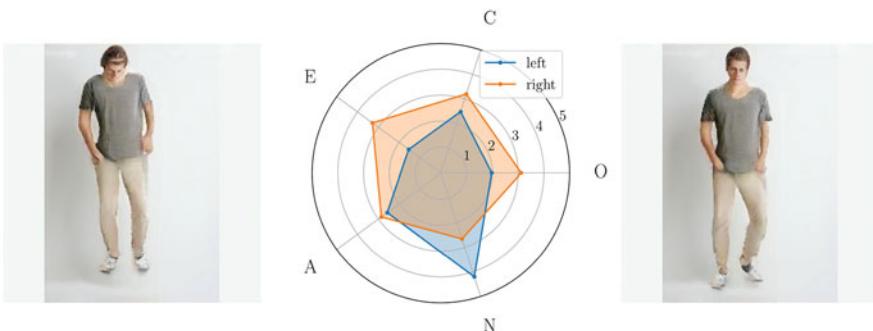


Fig. 17.7 Comparison of two sad poses

extroversion increase when the character faces the camera. The figure on the left has higher neuroticism and lower openness.

In Fig. 17.8, we compare two scared poses. Both figures keep their hands towards their heads, expressing scared gestures. The figure on the right is facing forward, while the one on the left is facing away from the camera. The knees of the right

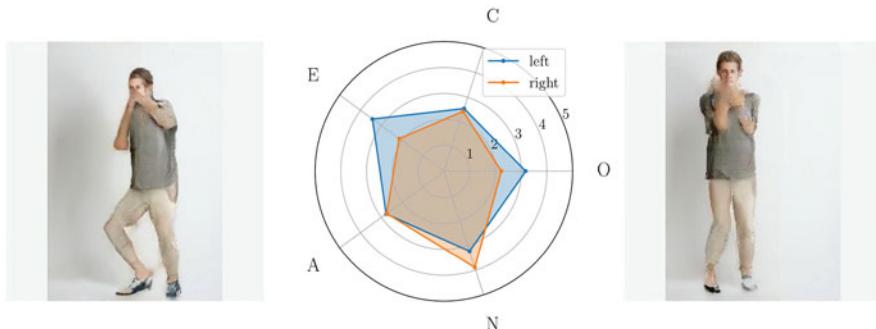


Fig. 17.8 Comparison of two scared poses

figure are spreading. In contrast, the figure on the right keeps his knees closer. We observe higher extroversion and openness in the left figure, probably due to the more relaxed posture. We believe such differences within the groups are the main reason for the mean scores being close to neutral. When the intensity of an expression is low, it can be mistaken for the opposite emotion. Grouping the poses based on emotional intensity can solve this issue.

17.5 Discussion

Facial expressions and full-body poses portray different personality traits based on the emotional categories they represent. Our experiments suggest that facial expressions can describe more diverse personality traits than body poses when presented in isolation. The broad recognition of emotions in basic facial expressions helps achieve stronger personality associations. In contrast, emotions in different body poses are culturally and contextually more varied, without universal connotations. Thus, subtle changes in pose can be misinterpreted by observers, resulting in mixed reactions with diminishing effects in terms of personality expression. These two modalities can be combined to complement each other for better recognition [68] and thus more precise personality expression [85]. In this respect, body poses can provide a context to facial expressions. A natural extension to this work would be to analyze the combination of emotional body poses and facial expressions. Aviezer et al. [9] have studied this by presenting juxtaposed photographs of extreme facial expressions onto bodies in various poses and contexts, such as a tennis player winning a match or a patient in pain. They found that the observers interpreted the same intense emotional facial expression either as joy or pain depending on the body pose. However, the semantics of the scene might have influenced the results, as they depicted different objects and environments such as a tennis racket or a hospital bed. Therefore, further research is needed to fully understand the influence of combining emotional body poses with facial expressions. Regarding the influence of context

on emotion recognition, we refer the readers to the chapter by Filntsis et al. [32] in this volume, where the authors explore the effect of different information streams on the automatic recognition of emotions.

The results of our two studies suggest that agreeableness is the best-represented personality factor in facial expressions, while extroversion is the best-represented factor in body poses. For facial expressions, emotions influence the perception of the personality factors in the same direction, suggesting a Big One [72]-like effect. For instance, happiness, a positive emotion, is associated with high openness, conscientiousness, extroversion, agreeableness, and emotional stability, all considered positive traits. In contrast, anger, a negative emotion, has low scores for all these traits. The Big One disposition is not observed in body poses, where emotions have varying effects on the perception of individual factors.

In general, we observe that emotional facial expressions can convey all the personality traits while emotional poses mostly control the perception of extroversion and openness. However, facial expressions also indicate a higher correlation between different personality factors. This is especially prominent in happiness; i.e., by employing a happy facial expression, we can express higher conscientiousness at the cost of a general personality shift in the positive direction. A conscientious yet disagreeable face needs more subtle control of AUs than the simple employment of a smile.

Following the quantitative LMA-based features used in skeletal animation [4, 7], we can form high-level descriptors for static 2D poses. For example, the area of the convex hull of a set of joints can be a metric related to LMA Space Effort, which measures the attention towards the surrounding space. The horizontal distance between the hands can also be a different metric for the same LMA parameter. One can devise many such features and construct a descriptive weighted linear combination, where the weight, or the importance, of each partial feature would be task-dependent. For example, Aristidou et al. [7] use Pearson correlation analysis to calculate the correlation between their interpretation of LMA features and the recognized emotion in short video clips. Partial weights of the linear combination can be adjusted to maximize the Pearson correlation with the subject of interest, similar to how a neural network trains. Our preliminary experiments show that a linear combination of horizontal distances between all joint pairs can achieve a Pearson correlation coefficient as high as 0.9 for extroversion in our emotional pose set. We leave forming a comprehensive LMA feature toolkit for static 2D poses as future work, which can be helpful in both recognition and synthesis tasks in affective computing.

17.6 Conclusion

We present our findings as a general guide in virtual human design for personality expression. Certainly, our test cases are highly broad and more research on the subtle details of facial expressions and body poses is needed to establish precise

connections between emotion expression and personality judgments. However, even by controlling the general aspects of emotional behavior, animators can customize the personality of virtual agents to enhance believability or improve communication. Emotional facial expressions and poses can be used together [85] or interchangeably based on scenario constraints. For example, emotional poses can be utilized in setups where the face is not prominent, and facial expressions can be preferred in close-up views.

Subtle pose elements are essential when the pose's emotional content is unclear. For instance, the distance of the hands to the body can influence the apparent extroversion. In this case, more precise control of the pose, for example, using LMA-based features [25, 85], can result in better personality expression. We publish our pose study data for further analysis. A closer look at the differences between the poses of the same emotional category can reveal exciting results. One possible future research direction is to evaluate whether compound emotions enhance or diminish certain effects, and if they have cultural associations that influence the perception of personality traits. Another research direction is to analyze the effect of adding contextual information to isolated facial expressions and poses.

References

1. Adams Jr, R.B., Nelson, A.J., Soto, J.A., Hess, U., Kleck, R.E.: Emotion in the neutral face: a mechanism for impression formation? *Cogn. Emotion* **26**(3), 431–441 (2012)
2. Adishesha, A.S., Zhao, T.: Emotion embedded pose generation. In: Proceedings of the European Conference on Computer Vision, ECCV '20, pp. 774–787. Springer (2020)
3. Ahmed, F., Bari, A.H., Gavrilova, M.L.: Emotion recognition from body movement. *IEEE Access* **8**, 11761–11781 (2019)
4. Ajili, I., Mallem, M., Didier, J.Y.: Human motions and emotions recognition inspired by LMA qualities. *Vis. Comput.* **35**(10), 1411–1426 (2019)
5. Allbeck, J., Badler, N.: Toward representing agent behaviors modified by personality and emotion. In: Proceedings of the Workshop on Embodied Conversational Agents at AAMAS '02, vol. 2 (2002)
6. Alsaidan, N., Menai, M.E.B.: A survey of state-of-the-art approaches for emotion recognition in text. *Knowl. Inf. Syst.* **62**(8), 2937–2987 (2020)
7. Aristidou, A., Charalambous, P., Chrysanthou, Y.: Emotion analysis and classification: understanding the performers' emotions using the LMA entities. *Comput. Graph. Forum* **34**(6), 262–276 (2015)
8. Aslan, S., Güdükbay, U., Dibeklioğlu, H.: Multimodal assessment of apparent personality using feature attention and error consistency constraint. *Image Vis. Comput.* **110**, 104163 (2021)
9. Aviezer, H., Trope, Y., Todorov, A.: Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* **338**(6111), 1225–1229 (2012)
10. Bänziger, T., Grandjean, D., Scherer, K.R.: Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT). *Emotion* **9**(5), 691 (2009)
11. Batrinca, L., Mana, N., Lepri, B., Sebe, N., Pianesi, F.: Multimodal personality recognition in collaborative goal-oriented tasks. *IEEE Trans. Multimedia* **18**(4), 659–673 (2016)
12. Beer, J.M., Fisk, A.D., Rogers, W.A.: Recognizing emotion in virtual agent, synthetic human, and human facial expressions. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, HFES '10, vol. 54, pp. 2388–2392. SAGE Publications, Los Angeles, CA (2010)

13. Biel, J.I., Teijeiro-Mosquera, L., Gatica-Perez, D.: FaceTube: predicting personality from facial expressions of emotion in online conversational video. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI '12, pp. 53–56 (2012)
14. Burton, S.J., Samadani, A.A., Gorbet, R., Kulić, D.: Laban movement analysis and affective movement generation for robots and other near-living creatures. In: Dance Notations and Robot Motion, pp. 25–48. Springer (2016)
15. Camurri, A., Lagerlöf, I., Volpe, G.: Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *Int. J. Hum. Comput. Stud.* **59**(1–2), 213–225 (2003)
16. Castellano, G., Kessous, L., Caridakis, G.: Emotion recognition through multiple modalities: face, body gesture, speech. In: Affect and Emotion in Human-Computer Interaction, pp. 92–103. Springer (2008)
17. Chi, D., Costa, M., Zhao, L., Badler, N.: The EMOTE model for effort and shape. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '00, pp. 173–182. ACM (2000)
18. Crane, E.A., Gross, M.M.: Effort-shape characteristics of emotion-related body movement. *J. Nonverbal Behav.* **37**(2), 91–105 (2013)
19. Dael, N., Mortillaro, M., Scherer, K.R.: Emotion expression in body action and posture. *Emotion* **12**(5), 1085 (2012)
20. Darwin, C., Prodgger, P.: *The Expression of the Emotions in Man and Animals*, 3rd edn. Oxford University Press, Oxford, UK (1998)
21. De Gelder, B., Van den Stock, J.: The bodily expressive action stimulus test (BEAST). Construction and validation of a stimulus basis for measuring perception of whole body expression of emotions. *Front. Psychol.* **2**, 181 (2011)
22. DeYoung, C.G., Hirsh, J.B., Shane, M.S., Papademetris, X., N.Rajeevan, R.Gray, J.: Testing predictions from personality neuroscience: brain structure and the big five. *Psychol. Sci.* **21**(16), 820–828 (2010)
23. Digman, J.M.: Personality structure: emergence of the five-factor model. *Annu. Rev. Psychol.* **41**(1), 417–440 (1990)
24. Durupinar, F., Güdükbay, U., Aman, A., Badler, N.I.: Psychological parameters for crowd simulation: from audiences to mobs. *IEEE Trans. Vis. Comput. Graph.* **22**(9), 2145–2159 (2015)
25. Durupinar, F., Kapadia, M., Deutsch, S., Neff, M., Badler, N.I.: PERFORM: Perceptual approach for adding OCEAN personality to human motion using Laban Movement Analysis. *ACM Trans. Graph.* **36**(1), Article no. 6, 16 pages (2016)
26. Eddine Bekhouche, S., Dornaika, F., Ouafi, A., Taleb-Ahmed, A.: Personality traits and job candidate screening via analyzing facial videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '17, pp. 10–13 (2017)
27. Ekmann, P.: Universal facial expressions in emotion. *Stud. Psychol.* **15**(2), 140–147 (1973)
28. Ekman, P.: Basic emotions. In: *Handbook of Cognition and Emotion*, vol. 98, no. 45–60, 16 (1999)
29. Ekman, P., Friesen, W.V., Hager, J.C.: *Facial Action Coding System*, 2nd edn. Research Nexus Division of Network Information Research Corporation, Salt Lake City, UT (2002)
30. Erkoç, Z., Demirci, S., Sonlu, S., Güdükbay, U.: Skeleton-based personality recognition using Laban movement analysis. In: *Understanding Social Behavior in Dyadic and Small Group Interactions*, Proceedings of Machine Learning Research, vol. 173, pp. 74–87. PMLR (2022)
31. Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.F., De Cock, M.: Computational personality recognition in social media. *User Model. User-Adapt. Interaction* **26**(2), 109–142 (2016)
32. Filmtidis, P.P., Efthymiou, N., Potamianos, G., Maragos, P.: Multi-stream temporal networks for emotion recognition in children and in the wild. In: Wang, J.Z., Reginald, J., Adams, B. (eds.) *Emotional Expression as a Means of Communicating Virtual Human Personalities*
33. Fridlund, A.J.: *Human Facial Expression: An Evolutionary View*. Academic Press (2014)

34. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., et al.: FD-GAN: Pose-guided feature distilling GAN for robust person re-identification. In: Proceedings of the 31th Advances in Neural Information Processing Systems, NeurIPS '18 (2018)
35. Glowinski, D., Camurri, A., Volpe, G., Dael, N., Scherer, K.: Technique for automatic emotion recognition by body gesture analysis. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW '08, pp. 1–6. IEEE (2008)
36. Gosling, S.D., Ko, S.J., Mannarelli, T., Morris, M.E.: A room with a cue: personality judgments based on offices and bedrooms. *J. Personal. Soc. Psychol.* **82**(3), 379 (2002)
37. Gosling, S.D., Rentfrow, P.J., Swann Jr, W.B.: A very brief measure of the big-five personality domains. *J. Res. Personal.* **37**(6), 504–528 (2003)
38. Groff, E.: Laban movement analysis: charting the ineffable domain of human movement. *J. Phys. Educ. Recreat. Dance* **66**(2), 27–30 (1995)
39. Gross, M.M., Crane, E.A., Fredrickson, B.L.: Methodology for assessing bodily expression of emotion. *J. Nonverbal Behav.* **34**(4), 223–248 (2010)
40. Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. *J. Netw. Comput. Appl.* **30**(4), 1334–1345 (2007)
41. Hortensius, R., Hekele, F., Cross, E.S.: The perception of emotion in artificial agents. *IEEE Trans. Cogn. Dev. Syst.* **10**(4), 852–864 (2018)
42. Hu, C., Walker, M.A., Neff, M., Tree, J.E.F.: Storytelling agents with personality and adaptivity. In: Proceedings of Intelligent Virtual Agents, pp. 181–193. Springer International Publishing (2015)
43. Hu, Y., Parde, C.J., Hill, M.Q., Mahmood, N., O'Toole, A.J.: First impressions of personality traits from body shapes. *Psychol. Sci.* **29**(12), 1969–1983 (2018)
44. Ioannou, S.V., Raouzaiou, A.T., Tzouvaras, V.A., Mailis, T.P., Karpouzis, K.C., Kollias, S.D.: Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Netw.* **18**(4), 423–435 (2005)
45. Isbister, K., Nass, C.: Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *Int. J. Hum. Comput. Stud.* **53**(2), 251–267 (2000)
46. Jack, R.E., Garrod, O.G., Yu, H., Caldara, R., Schyns, P.G.: Facial expressions of emotion are not culturally universal. *Proc. Natl. Acad. Sci.* **109**(19), 7241–7244 (2012)
47. Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., Zareapoor, M.: Hybrid deep neural networks for face emotion recognition. *Pattern Recogn. Lett.* **115**, 101–106 (2018)
48. John, O. P. and Naumann, L. P., and Soto, C. J.: Paradigm shift to the integrative big five trait taxonomy: history, measurement, and conceptual issues. In: John, O.P., Robins, R.W., Pervin, L.A. (eds.) *Handbook of Personality. Theory and Research*, 3rd edn., pp. 114–158. Guilford Press, New York, NY (2000)
49. Jolliffe, D., Farrington, D.P.: Development and validation of the basic empathy scale. *J. Adolescence* **29**(4), 589–611 (2006)
50. Kaushal, V., Patwardhan, M.: Emerging trends in personality identification using online social networks—a literature survey. *ACM Trans. Knowl. Discov. Data* **12**(2), 1–30 (2018)
51. Khosdelazad, S., Jorna, L.S., McDonald, S., Rakers, S.E., Huitema, R.B., Buunk, A.M., Spikman, J.M.: Comparing static and dynamic emotion recognition tests: performance of healthy participants. *Plos One* **15**(10), e0241297 (2020)
52. Kilts, C.D., Egan, G., Gideon, D.A., Ely, T.D., Hoffman, J.M.: Dissociable neural pathways are involved in the recognition of emotion in static and dynamic facial expressions. *Neuroimage* **18**(1), 156–168 (2003)
53. Kim, H., Kwak, S.S., Kim, M.: Personality design of sociable robots by control of gesture design factors. In: Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, ROMAN '08, pp. 494–499. IEEE (2008)
54. Ko, B.C.: A brief review of facial emotion recognition based on visual information. *Sensors* **18**(2), 401 (2018)
55. Komulainen, E., Meskanen, K., Lipsanen, J., Lahti, J.M., Jylhä, P., Melartin, T., Wichers, M., Isometsä, E., Ekelund, J.: The effect of personality on daily life emotional processes. *PloS One* **9**(10), Article no. e110907, 9 pages (2014)

56. Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech: a review. *Int. J. Speech Technol.* **15**(2), 99–117 (2012)
57. Koppenstein, M.: Motion cues that make an impression: predicting perceived personality by minimal motion information. *J. Exp. Soc. Psychol.* **49**(6), 1137–1143 (2013)
58. Likert, R.: A technique for the measurement of attitudes. *Arch. Psychol.* **140**, 5–55 (1932)
59. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR '16 (2016)
60. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping GAN: a unified framework for human motion imitation, appearance transfer and novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, CVPR '19, pp. 5904–5913 (2019)
61. Longua, J., DeHart, T., Tennen, H., Armeli, S.: Personality moderates the interaction between positive and negative daily events predicting negative affect and stress. *J. Res. Personal.* **43**(4), 547–555 (2009)
62. Lopez, L.D., Reschke, P.J., Knothe, J.M., Walle, E.A.: Postural communication of emotion: perception of distinct poses of five discrete emotions. *Front. Psychol.* **8**, 710 (2017)
63. Lucas, R.E., Baird, B.M.: Extraversion and emotional reactivity. *J. Personal. Soc. Psychol.* **86**(3), 473 (2004)
64. Luo, Y., Ye, J., Adams Jr, R.B., Li, J., Newman, M.G., Wang, J.Z.: ARBEE: towards automated recognition of bodily expression of emotion in the wild. *Int. J. Comput. Vis.* **128**(1), 1–25 (2020)
65. Ma, L., Deng, Z.: Real-time facial expression transformation for monocular RGB video. *Comput. Graph. Forum* **38**(1), 470–481 (2019)
66. Mairesse, F., Walker, M.: PERSONAGE: personality generation for dialogue. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, ACL '07, pp. 496–503 (2007)
67. Masuda, M., Kato, S.: Motion rendering system for emotion expression of human form robots based on laban movement analysis. In: Proceedings of the 19th International Symposium in Robot and Human Interactive Communication, ROMAN '10, pp. 324–329. IEEE (2010)
68. Meeren, H.K., van Heijnsbergen, C.C., de Gelder, B.: Rapid perceptual integration of facial expression and emotional body language. *Proc. Natl. Acad. Sci.* **102**(45), 16518–16523 (2005)
69. Miller, D.J., Vachon, D.D., Lynam, D.R.: Neuroticism, negative affect, and negative affect instability: establishing convergent and discriminant validity using ecological momentary assessment. *Personal. Individ. Differences* **47**(8), 873–877 (2009)
70. Moon, Y., Nass, C.: How “real” are computer personalities? *Commun. Res.* **23**(6), 651–674 (1996)
71. Moreno-Armendáriz, M.A., Martínez, C.A.D., Calvo, H., Moreno-Sotelo, M.: Estimation of personality traits from portrait pictures using the five-factor model. *IEEE Access* **8**, 201649–201665 (2020)
72. Musek, J.: A general factor of personality: evidence for the big one in the five-factor model. *J. Res. Personal.* **41**(6), 1213–1233 (2007)
73. Neff, M., Wang, Y., Abbott, R., Walker, M.: Evaluating the effect of gesture and language on personality perception in conversational agents. In: Proceedings of the International Conference on Intelligent Virtual Agents, IVA '10, pp. 222–235. Springer (2010)
74. Pelachaud, C.: Modelling multimodal expression of emotion in a virtual agent. *Philos. Trans. R. Soc. B Biol. Sci.* **364**(1535), 3539–3548 (2009)
75. Polzehl, T., Möller, S., Metze, F.: Automatically assessing personality from speech. In: Proceedings of the IEEE Fourth International Conference on Semantic Computing, ICSC '10, pp. 134–140. IEEE (2010)
76. Randhavane, T., Bhattacharya, U., Kapsakis, K., Gray, K., Bera, A., Manocha, D.: Identifying emotions from walking using affective and deep features. Preprint. arXiv:1906.11884 (2019)

77. Riggio, H.R., Riggio, R.E.: Emotional expressiveness, extraversion, and neuroticism: a meta-analysis. *J. Nonverbal Behav.* **26**(4), 195–218 (2002)
78. Roberts, B., DelVecchio, W.: The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. *Psychol. Bull.* **126**, 3–25 (2000)
79. Rushton, J.P., Irving, P.: A general factor of personality (GFP) from the multidimensional personality questionnaire. *Personal. Individ. Differences* **47**(6), 571–576 (2009)
80. Scheff, T.: Toward defining basic emotions. *Qualitat. Inquiry* **21**(2), 111–121 (2015)
81. Scherer, K.R., Ellgring, H., Dieckmann, A., Unfried, M., Mortillaro, M.: Dynamic facial expression of emotion and observer inference. *Front. Psychol.*, pp. Article no. 508, 17 pages (2019)
82. Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., Yang, X.: A review of emotion recognition using physiological signals. *Sensors* **18**(7), 2074 (2018)
83. Siarohin, A., Lathuilière, S., Sangineto, E., Sebe, N.: Appearance and pose-conditioned human image generation using deformable GANs. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(4), 1156–1171 (2019)
84. Smith, H.J., Neff, M.: Understanding the impact of animated gesture performance on personality perceptions. *ACM Trans. Graph.* **36**(4), Article no. 59, 12 pages (2017)
85. Sonlu, S., Güdükbay, U., Durupinar, F.: A conversational agent framework with multi-modal personality expression. *ACM Trans. Graph.* **40**(1), Article no. 7, 16 pages (2021)
86. Subramanian, R., Wache, J., Abadi, M.K., Vieriu, R.L., Winkler, S., Sebe, N.: ASCERTAIN: emotion and personality recognition using commercial sensors. *IEEE Trans. Affect. Comput.* **9**(2), 147–160 (2016)
87. Tan, S.C., Nareyek, A.: Integrating facial, gesture, and posture emotion expression for a 3D virtual agent. In: Proceedings of the 14th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational and Serious Games, CGames 2009, pp. 23–31 (2009)
88. Tang, H., Wang, W., Wu, S., Chen, X., Xu, D., Sebe, N., Yan, Y.: Expression conditional GAN for facial expression-to-expression translation. In: Proceedings of the IEEE International Conference on Image Processing, ICIP '19, pp. 4449–4453. IEEE (2019)
89. Timothy J, T., Widiger, T.A.: Dimensional models of personality: the five-factor model and the DSM 5. *Dialogues Clin. Neurosci.* **15**(2), 135–146 (2013)
90. Tinwell, A., Grimshaw, M., Nabi, D.A., Williams, A.: Facial expression of emotion and perception of the uncanny valley in virtual characters. *Comput. Hum. Behav.* **27**(2), 741–749 (2011)
91. Trnka, R., Tavel, P., Hašto, J.: Facial expression of fear in the context of human ethology: recognition advantage in the perception of male faces. *Neuroendocrinol. Lett.* **36**(2), 106–111 (2015)
92. Tukey, J.W.: Comparing individual means in the analysis of variance. *Biometrics* **5**(2), 99–114 (1949)
93. Tzirakis, P., Trigeorgis, G., Nicolaou, M.A., Schuller, B.W., Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process.* **11**(8), 1301–1309 (2017)
94. Wang, Y., Tree, J.E.F., Walker, M., Neff, M.: Assessing the impact of hand motion on virtual character personality. *ACM Trans. Appl. Perception* **13**(2), 1–23 (2016)
95. Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., Kissler, J.: Mapping the emotional face. how individual face parts contribute to successful emotion recognition. *PloS One* **12**(5), Article no. e0177239, 15 pages (2017)
96. Williams, A.C.d.C.: Facial expression of pain: an evolutionary account. *Behav. Brain Sci.* **25**(4), 439–455 (2002)
97. Wolf, K.: Measuring facial expression of emotion. *Dialogues Clin. Neurosci.* **17**(4), 457–462 (2022)
98. Xu, Q., Yang, Y., Tan, Q., Zhang, L.: Facial expressions in context: electrophysiological correlates of the emotional congruency of facial expressions and background scenes. *Front. Psychol.* **8**, 2175 (2017)

99. Yamagata, S., Suzuki, A., Ando, J., Ono, Y., Kijima, N., Yoshimura, K., Ostendorf, F., Angleitner, A., Riemann, R., Spinath, F.M., Livesley, W.J., Jang, K.L.: Is the genetic structure of human personality universal? A cross-cultural twin study from North America, Europe, and Asia. *J. Personal. Soc. Psychol.* **90**(6), 987–998 (2006)
100. Yang, B., Fotios, S.: Lighting and recognition of emotion conveyed by facial expressions. *Light. Res. Technol.* **47**(8), 964–975 (2015)

Chapter 18

Modeling Emotion Perception from Body Movements for Human-Machine Interactions Using Laban Movement Analysis



Tal Shafir

Abstract Emotion recognition from body movements is based in humans on the existence of associations in our brains between certain movements and specific emotions. These associations are responsible for the expression of emotions using typical movements, and for the elicitation of emotions by moving those typical associated movements. When observing other people move, our mirror neurons simulate in our brains the movements we see. This simulation elicits the emotion associated with the observed movements, enabling us to recognize it. To enable AI to recognize through computer vision the emotions expressed in movements, we have to provide the computer information about which movements are associated with which emotion. This chapter describes research which identified these associations, using Laban Movement Analysis (LMA) to portray the movements. We first demonstrated that both execution and observation of movements that express anger, happiness, sadness or fear, elicit the associated emotion. Next, we extracted from the movements used in the first experiment the LMA motor components that constructed those movements. Examining the effects of moving different combinations of those components, we identified which motor components are associated with, and enhance which emotion. We then established that movements composed of motor components associated with a specific emotion are perceived as expressing that emotion. Lastly, we demonstrated automatic recognition of these motor components using machine learning. Using LMA motor components (motor characteristics) to characterize emotional motor behavior rather than using a list of specific movements enables the identification of the associated emotion in any motor behavior.

T. Shafir (✉)

The Emili Sagol Creative Arts Therapies Research Center, University of Haifa, Haifa, Israel

Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA

e-mail: gahl@umich.edu

18.1 Introduction

The accelerated development in recent years of technologies like robotics, artificial intelligence (AI), virtual reality (VR), and computer vision, causes human-machine interactions to become more and more prevalent, and soon they will become part of everyone's daily life. Many tasks which are currently carried out by human-beings, from serving customers in various types of businesses, to medical and psychiatric diagnosis, care taking of disabled or elderly people, driving, and more, will soon, if not already, be performed by computers, robots, and various devices using computer vision, sensors and AI. In addition, VR is increasingly being used for various trainings, education, entertainment, and more. In order for these human-machine interactions to be successful and attain their goals, they necessitate good communication between the human and the machine, which requires the machine to correctly identify, understand and predict the human's behavior: motivation, decisions, intentions and actions. Only such correct identification will enable the machine to communicate with people effectively and to respond to humans' behavior appropriately. Because humans are driven and activated by their emotions [21, 26, 34, 38], correct emotion recognition by the machine is fundamental for these interactions.

Humans express their emotions through their behavior, in particular their vocal expressions, facial expressions, and whole-body expressive movements and posture. While people not always use their voice to express themselves, and while facial expressions are much smaller in size, and are not always visible (depending on where the face is directed), whole-body expressions are often bigger and can be captured by a camera from further away and from any direction. In addition, while people are usually more aware of their facial and vocal emotional expressions and learn to control or hide them, most people are less conscious of their whole-body emotional expressions. Thus, whole-body expressions usually provide the most accurate indication to people's authentic emotions. Automatic emotion recognition from whole-body movement and posture is therefore of crucial importance for machines when interacting with humans, and raises the question of how best to model it for the purpose of human-machine interaction. To answer this question, we first have to understand what is emotion, and how is it expressed through the body.

18.2 Emotion and Its Relation to the Body

The biological mechanism for emotion recognition from body movements is based in humans on activation of the mirror neurons, and the existence of associations in our brain between certain movements and specific emotions. There have been many definitions to 'emotion' along history and in different fields of study and for the purpose of this chapter I will use the definition of the famous neurologist and neuroscientist Antonio Damasio. According to Damasio, emotions are "part

of a multitiered and evolutionarily set neural mechanism, aimed at maintaining organism homeostasis. This mechanism is based on brain structures that regulate the organism's current state by executing specific actions via the musculoskeletal system, ranging from facial and postural expression to complex behaviors, and by producing chemical and neural responses aimed at the internal milieu, viscera and telencephalic neural circuits" [16, p. 1049]. According to Damasio, the current state of the organism's (human) body is conveyed to the brain through proprioception (input to the brain from the muscles and joints, which gives the brain information about posture and movement) and through interoception (input from the body to the brain representing the physiological state of the body, such as thermal, metabolic, hormonal state, etc.). These inputs from the body to the brain create unique neural activation patterns in the brain, whose purpose is to keep us alive by causing us to behave in a way that will maintain our homeostasis. These neural activation patterns represent, according to Damasio, unconscious emotions that guide behavior and influence decisions, and they correlate with the conscious feelings of those emotions, feelings which we consciously experience as fear, anger, happiness etc. [14–16].

The meaning of Damasio's definition of 'emotion' is that certain interoceptive and proprioceptive inputs from the body are associated in the brain with specific emotions (and that is why and how they elicit the correlated feelings), and the implication of this proposition is that by controlling and changing our motor behavior we can change the interoceptive and proprioceptive inputs to the brain and consequently change and regulate our emotions and feelings. This idea that by using certain motor behaviors we can enhance or elicit specific feelings was demonstrated in studies both with facial expressions (e.g., [9]) and bodily movements and postures (e.g., [50, 56]).

In addition to this neuroscientific assertion that feedback from the body to the brain creates and affects our emotions and feelings, another important neuroscientific notion is that of the mirror neurons. Mirror neurons are neurons in the brain which are activated very similarly during execution of movements and during observation and imagination of the same movements. The activation of the mirror neurons during observation and imagination of movements is assumed to simulate the brain's motor commands for the execution of the same movements, as well as to simulate the expected sensory, interoceptive and proprioceptive afferent input from the body to the brain which happens when performing those movements. This simulation is hypothesized to enhance or induce the same emotions and feelings as those enhanced by the actual execution of those movements and their resulted actual input from the body to the brain [25, 30, 44, 53, 63]. This enhancement of specific emotions and feelings following observation of expressive movements, is the base for our ability to perceive and recognize the mover's feelings, and our ability to empathize with them and understand their intentions, when watching their movements.

In the computer science disciplines, such as computer vision, robotics, artificial intelligence, etc., researchers do not differentiate between emotions and feelings and usually use the term 'emotion' to describe the conscious experience that

Damasio defined as feeling. Thus, to be in line with the professional literature and terminology used in these fields, I will use in the rest of this chapter the term ‘emotion’ to describe what Damasio defined as feeling.

18.3 How to Model Expressive Movements

As mentioned before, the biological mechanism for emotion recognition in humans is based on the fact that using our mirror neurons we simulate in our brains the movement that we see. The simulation of expressive movement enhances within us the emotion that is associated in our brain with that movement. Feeling this emotion enables us to recognize what was the emotion that the mover expressed, since both observation and execution of the same movements enhance the same emotion. In other words, observing a person expressing an emotion with his body causes us to feel the same emotion that he expresses.

When it comes to emotion recognition by machines during human-machine interaction, and using computer vision methodology, this biological mechanism which works automatically for humans cannot be used, because machines cannot feel. Nevertheless, if we know which movements are associated with each emotion, and if we can use computer vision to capture and identify those movements, then the machine will be able to correctly identify its human counterpart’s emotions.

Many researchers in the field of psychology have tried to characterize the movements associated with various emotions during their bodily expressions and/or during their perception from body movements and posture. Some researchers, such as Darwin [17], Wallbott [60] or Dael [13], identified several specific movements executed with specific body parts, which were associated with specific emotions. For example, head hangs on contracted chest was found as characterizing sadness [17], lifting the shoulders was found as typical to elated joy or hot anger [60], and both arms at rest and in the pockets were associated with sadness or relief [13]. The main limitation of these studies is that their findings are based on a limited number of movements performed by a limited number of people who participated in those studies. As such, these movements do not cover the entire possible spectrum of emotional expressions. This limitation is intensified by the fact that many of these studies were based on exaggerated stereotypical emotional expressions performed by actors (e.g., [13, 60], and they didn’t take into consideration, nor did they examine, the authentic and more refined bodily expressions which are usually observed in the wild. Moreover, gender and cultural differences in emotional expressions were also not considered. Thus, the list of emotional movements associated with each emotion based on these studies is incomplete.

To overcome this limitation, we decided to explore not which specific movements are associated in the brain with which specific emotions, but which movement characteristics are associated with each emotion. Defining emotional expressions based on their motor characteristics as opposed to specific movements, enables to

characterize, model and recognize the emotion expressed in any bodily emotional expression, even those which were not observed in previous studies.

Other researchers also used coding systems that included movement characteristics as opposed to specific movements. Some used various movement dimensions. For example, De Meijer used movements in the vertical and sagittal direction, force, velocity, and directness [18] to characterize expressive movements, while Montepare used form, tempo, force and direction [43]. A few researchers used Laban Movement Analysis (LMA) or its most known components: Effort and Shape [10, 27, 28, 39]. Yet others characterized the movements based on the specific muscles that are activated during the expressive movements of each emotion [31, 32], or used kinematic variables such as movement time, velocity, acceleration, joints displacement (range of motion), and joints coordination [5, 27, 28, 48, 52, 54]. For review of these studies and summary of the different movement characteristics that were found to be associated with each emotion see [36, 62].

Although these studies were able to discriminate among the different emotions expressed in movement, they used different coding schemes, making it very difficult to compare outcomes across studies and to build a comprehensive description of the associations between certain motor characteristics of body movements and specific emotions [28].

To overcome these difficulties, we chose to describe and characterize the movements in our studies using Laban Movement Analysis (LMA), which, to the best of our knowledge, is the most comprehensive movement analysis system that exists. Analyzing movements using LMA is advantageous over other methods, as it captures both qualitative motor components (movement characteristics) and quantitative aspects of the movement, which can be easily described by features that depict the movement kinematics and/or dynamics.

LMA's comprehensiveness as a method for analyzing movements could be inferred from its diverse use in research: it has been used to evaluate fighting behaviors of rats [22], to analyze behavior of nonhuman animals in naturalistic settings [20], to diagnose autistic individuals [47], to evaluate motor recovery of stroke patients [23], to characterize the development of infants' reaching movements [24], to analyze and classify the movements and emotional states during dance performance [3, 61], or theater [55], and more. Many studies have also used it to describe, identify or create bodily emotional expressions for applications in human-robot interactions, interactive games such as the Xbox, and in animations e.g., [8, 12, 33, 40, 41, 51, 65, 66], and it has even been attempted, through the use of electroencephalography (EEG), to identify the brain mechanisms underlying the production of some of the LMA motor components [11].

Additional weakness of many of these studies (even those which used LMA to characterize the bodily emotional expressions), both in the psychological and computer science fields, was that they examined only a small list of movement characteristics which they hypothesized (but without confirming those hypotheses through research) to be related to emotional expressions, extracting them from only a small number of emotional expressive movements. For example, Out of over 100 different Laban motor components, Zacharatos et al. [65] used only

two: the Space Effort and Time Effort, to recognize and differentiate between the four emotional states: concentration, meditation, excitement and frustration, using as data only 197 two-seconds clips of expressive movements performed by only 13 different people, and Ono et al. [46] used only three motor components: The Time, Space and Weight Efforts to recognize eight emotional states: surprise, concentration, pleasant, contented, bored, unpleasant and nervous, using only 1140 video frames (11.4 seconds) of motor expressions of each participant, with only 20 participants. However, human movement is a very complex phenomenon that can be described by numerous features (for examples of some of these numerous possible features see [1, 37]) and many studies would have been required to examine and find out which of those countless features are relevant and necessary for accurate emotion recognition from bodily expressions of different emotions. To overcome this difficulty, we used a different approach: We decided to examine first which Laban motor components are associated with which emotion. Once we know which movement characteristics are associated with each emotion, we can teach the computer to recognize people's emotional state by looking for the presence of those specific motor components in the people's movements. Such motor component identification can be done from any movement, enabling to recognize emotional expressions even from movements that the computer has never seen before. We believe that this strategy is the most effective and most efficient strategy for finding the movement features which are needed for computer perception of emotion from any bodily expressions.

Before I describe our experiments and results, I would like to give a brief review of LMA and its motor components, so that the results will make more sense.

18.4 Laban Movement Analysis (LMA)

LMA identifies 4 major movement categories: Body, Effort, Shape and Space, and each category has several subsets of motor components (LMA terms are spelled with capital letters to differentiate them from regular usage of these words.)

The Body category describes what is moving and it is composed of the components: Body segments (e.g., arm, legs, head), their coordination, and basic Body Actions, like Locomotion, Jump, Rotation, Change of Support, etc.

The Effort category describes the qualitative aspect of movement, or how we move. It expresses a person's inner attitude toward the movement, and it has four main factors, each describing the continuum between two extremes: indulging in the motor quality of that factor and fighting against that quality. The four factors of Effort are:

- (1) The Weight Effort represents the amount of force or pressure exerted by the body. Weight Effort can be Strong (using a lot of force), Light (using little force), or there might be a lack of weight/force activation, when we give in to the pull of gravity. In this case we say that there is Passive or Heavy Weight;

- (2) The Space Effort denotes the attention to the space/environment in which the movement is performed and the focus, or attitude toward a chosen pathway, i.e., is the movement Direct (i.e., the attention while moving is clearly aimed towards a specific point in space) or Flexible/Indirect;
- (3) Time Effort describes the mover's attitude towards time, or the degree of urgency. This effort is often manifested in the amount of acceleration involved in the movement, i.e., is the movement Sudden (with high acceleration) or Sustained (with constant velocity);
- (4) Flow Effort describes the element of control or the degree to which a movement is Bound, i.e., restrained or controlled by muscle contraction (usually co-contraction of agonist and antagonist muscles), vs. Free, i.e., being released and liberated.

The Space category describes where the movement goes in the environment. It describes many spatial factors such as the Direction (i.e., where the movement goes in space), which can be: Up or Down in the vertical dimension, Side open or Side across in the horizontal dimension and Forward or Backward in the sagittal dimension; the Level of the movement in space relative to the entire body or parts of the body, and can be Low level (movement towards the ground), Middle level, (movement maintaining level, without lowering or elevating) or High level (moving upward in space); Paths or how we travel through space by locomoting; and Pathways of body parts inside and through the Kinesphere. The Kinesphere is the personal bubble of reach-space around each mover, within and through which one can move without locomoting or traveling. Movement in the Kinesphere might take Central pathways, crossing the personal space close to the mover's center of the body, Peripheral pathways along the periphery of the mover's reach space, or Transverse pathways cutting across the reach space.

The fourth LMA category, Shape, reflects why we move: Shape describes how the body adapts its shape as we respond to our internal needs or to the environment: Do I want to connect with or avoid something, dominate or cower under? The way the body sculpts itself in space reflects relationship to self, others or to the environment. This category includes Shape flow which describes how the body changes to relate to oneself, i.e., to adapt to one's needs; it includes Shape change which describes changes in the form or shape of the body and includes the motor components of Expanding the body or Condensing it in all directions, and Rising or Sinking in the vertical dimension, Spreading or Enclosing in the horizontal dimension and Advancing or Retreating in the sagittal dimension. Another Shape component is Shaping and Carving which describes how a person shapes their body to adapt to, or to shape and affect the environment or other people. For example, when we hug someone, we might shape and carve our body, adjusting it to the shape of the other person's body.

In addition, another important aspect of LMA which is particularly helpful and meaningful to expression, is the Phrasing of movements. Phrasing describes changes over time, such as changes in the intensity of the movement over time, and similar to musical phrases, a movement phrase can Increase, Decrease, be Rhythmic, and

more. Phrasing can also depict how movement components shift during the same action or a series of actions occurring over time, for example beginning forcefully with Strong Weight then ending by making a Light Direct point conclusion.

18.5 Which Laban Motor Components Are Associated with Which Emotion

To uncover which Laban motor components are associated in our brain with each of the four basic emotions: anger, fear, sadness and happiness, we performed a series of studies:

In the first study [57], we aimed to demonstrate that motor execution, motor observation and motor imagination of emotional/expressive movements will enhance their associated feelings, and that these associations between certain movements and specific emotions will be reflected in different brain activation patterns when observing movements that express different emotions. To achieve this goal, we used video clips of whole-body emotional expressions of sadness, fear and happiness that were created and validated by Atkinson [4]. We taught the study's participants to move the same movements as in those clips (without revealing the purpose of the study), and asked them to move those movements for the motor execution task, to imagine themselves moving those movements as the motor imagination task, and to observe those clips for the motor observation task. We asked the participants to rate their emotions before and after moving (motor execution), observing and imagining the movements in those clips, and measured the difference in the rated emotional intensity between before and after each task for each emotion. We also asked the participants to observe those movements while their brain was scanned using fMRI (functional Magnetic Resonance Imaging). The results of the behavioral study showed that motor execution, observation and imagination of emotional movements enhanced the subjective feelings and affective state in the corresponding direction, i.e., moving, observing and imagining sad movements enhanced feelings of sadness, moving, observing and imagining fearful movements enhanced feelings of fear, etc. [57]. In addition, the fMRI results demonstrated that some regions of the brain which are responsible in general to the task of observing emotional movements (e.g., the Occipital cortex which is related to visual processing in general, and the insula and thalamus which are associated with emotional processing in general—all marked in the figure with black frames and arrows pointing to their location in the brain), were activated in all emotional observation conditions (Fig. 18.1). However, the activation of other regions was specific to observation of only emotional movements expressing a specific emotion: Observing only the happy movements but not observing the fearful or sad movements, activated the ventral striatum (marked in the figure with a red frame), which is a brain area involved in feelings of reward and motivation. Only movements expressing fear, when watching them, increased activation in

the parietal lobe and a much larger area, compared to the activation during happy movements, in the motor cortex (marked with green frames). The parietal lobe has to do with orienting in space, and together with increased activation in the motor cortex, the activation of these areas was probably related to getting ready to fight or fly, when we need to figure out where in the space around us is our enemy located, and where can we run to, in order to fly from the danger. Lastly, watching sad movements caused minimal activation in the brain as a whole, which goes along with the feeling of emptiness and not being able to do anything when being very sad (Fig. 18.1).

Once we demonstrated that moving and observing the emotional movements shown in Atkinson's clips can affect people's feelings in the corresponding direction, in our next study we aimed to uncover which LMA motor components are associated with each of the four basic emotions: anger, fear, sadness and happiness. To do that we first identified for each emotion the LMA motor components that appeared in the majority of Atkinson's clips of motor expressions of that emotion. We then created from these components motifs which are written instructions for movement, similar to scores for music. Each motif included a different combination

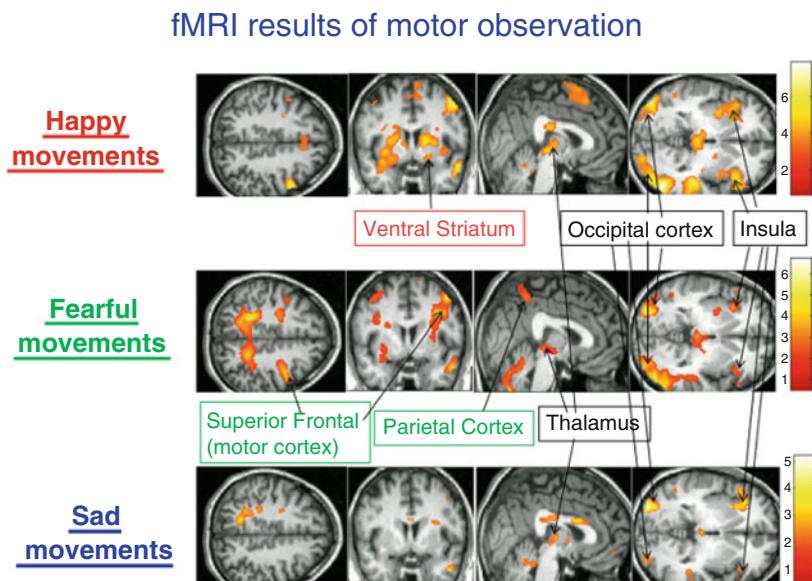


Fig. 18.1 fMRI results showing brain activation during observation of emotional movements. The upper row shows slices of the brain during observation of happy movements, the middle row shows the exact same brain slices during observation of fearful movements and the lower row during observation of sad movements. The side bar to the right of each row serves to indicate the amount of brain activation, which increases as the color changes from red to yellow. As can be seen, some regions of the brain (marked with black frames) were activated during observation of all emotional movements, while others (marked with red or green frames) were activated only during observation of a specific emotion. For further explanation see the text

of components that originally belonged to clips of the same emotion (Fig. 18.2). We then asked Laban experts to read the motifs, move them for a few minutes and rate what emotion they felt as a result of the movement. Reading the motifs and moving them is similar to reading musical score and singing or playing the notes in it. Because the instructions that we gave were written, and included only motor components, as opposed to showing them videos of specific movements, each of the Laban expert could execute the motif by doing their own specific movements which were different from person to person, even though they all included the same motor components that we instructed them to move. An example to clarify this point: the combination of the motor components Strong, Direct, Sudden and Forward (motif 13.12) could be moved as a punch forward with one fist, as punching with two fists forward simultaneously or one arm after another, as a sharp Karate like strike forward with the edge of the hand, as a fast kick forward, etc. All of these are completely different movements, using different body parts, but they all include the motor characteristics of moving fast (the Sudden component), with a lot of force (the Strong component) directed at a specific point in space (the Direct component), which is located in front of the person (the Forward component).

Eighty different experts participated in the study and all together we had results from 1241 executed motifs. Using statistical analysis, we predicted which motor components when moved, enhanced which emotion: Anger was enhanced by movements which included the motor components Strong Weight Effort, Sudden Time Effort, Direct Space Effort and Advance (a Shape component); Retreating, Condensing, movements with Bound Flow Effort, moving Backward, and Enclosing enhanced feeling fear; Passive/Heavy movements, bringing the arms to touch the upper body (chest, shoulder(s), head or face). Sinking in the torso and dropping the head down enhanced sadness; And happiness was elicited by the Laban motor components: Jump, Rise, Spread, Free Flow, Light Weight, Upward and Rhythmic Movements [58, 59].

To strengthen the idea of associations in the brain between certain movement characteristics and specific emotions, we wanted to show that the motor characteristics that moving them enhance a certain emotion, are the same motor characteristics found in movements that express that emotion.

To do that we asked the participants of a different study to recall a happy life event, to write in a short paragraph what happened during that event, to tell it to the experimenter, and then to move in the same way as they moved during that event or to express in movement what they felt during that event, or what they feel now when they recall that event. Then they were asked to rate the intensity of their emotion while they moved. These procedures were then repeated with recalling a sad life event. We filmed the participants during this entire process and certified Laban experts coded the motor components that appeared in their movements both during the conversation with the experimenter and during the motor expression. We then measured which motor components were most frequent in the movements that express happiness and which were most frequent in the movements expressing sadness.

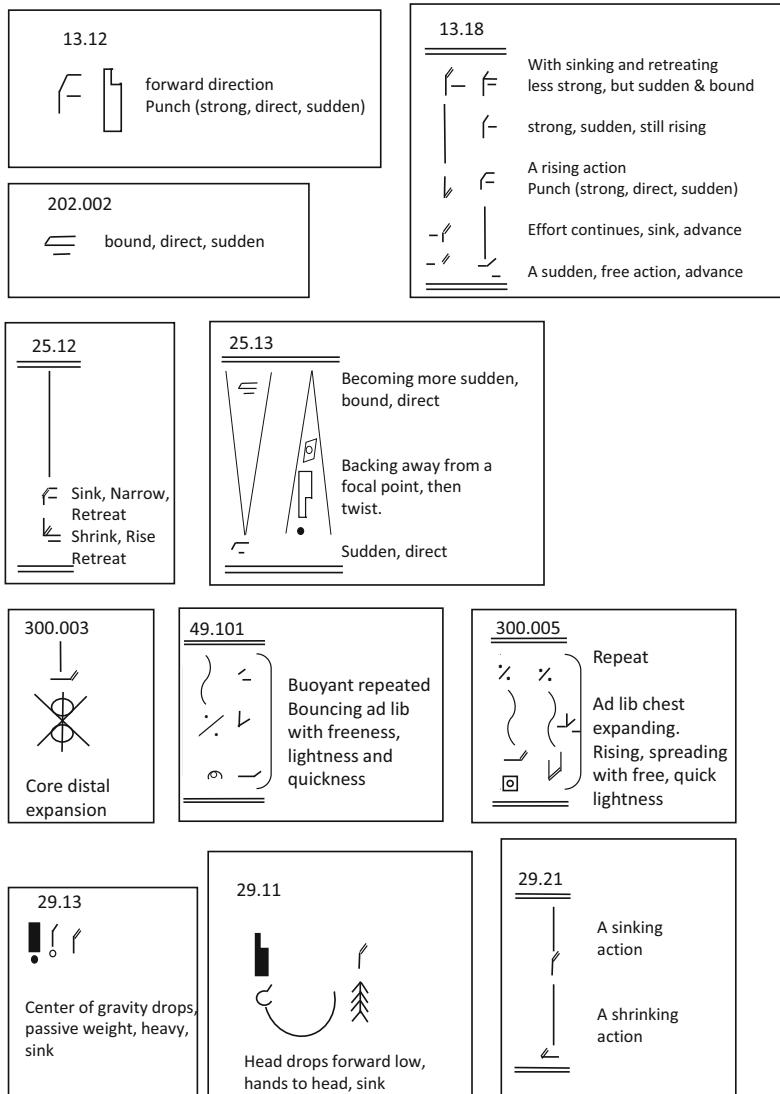


Fig. 18.2 Examples of motifs that the Laban experts read and moved. (Taken from [58]). Motifs 13.12, 202.002 and 13.18 are constructed from motor components that were taken from angry clips; motifs 25.12 and 25.13 include motor components taken from movements expressing fear; motifs 300.003, 49.101 and 300.05 were constructed from motor components from happy clips, and motifs 29.13, 29.11 and 29.21 include motor components from sad clips

The results showed that the happiness expressions were characterized by the same motor components as the ones we found in the previous experiment plus two new components: Sudden Time Effort and Rotations. The sadness motor expressions

were also characterized by the same movement components that were found in the previous experiment plus two new components: moving in a small personal Kinesphere, which means basically moving small movements, and Stillness, which means not moving at all.

The findings of additional motor components can be explained by the fact that this study was based on movements of 60 different people while Atkinson's clips from which we extracted the original motor components were based on the movements of only a handful of actors, and it is possible that these few actors did not use in their movements all of the emotional movement characteristics which exist for each emotion. Another important finding of this study was that the motor component of Sudden was associated with happiness, while in our previous study it was associated with anger. Happiness and anger are both considered to be "approach" emotions as opposed to fear and sadness which are emotions related to withdrawal or avoidance. I now believe that some of the motor components might be associated not with a discrete emotion but with a certain emotional aspect, and the exact emotion that will be associated to a motor expression depends on the specific combination of motor components, so that certain components can be associated with different emotions depending on which other motor components they are combined with.

We also found positive correlation between the frequency of appearance of the associated motor components and the emotional intensity, meaning that the more frequently the person used a certain motor component in his movement, the more intense was the feeling of the associated emotion.

Lastly, to demonstrate that when people observe movements which include those motor components, they recognize the movements as expressing the associated emotion, we created 113 three-seconds clips, each of which depicted a Laban expert performing a combination of 2–4 motor components which are associate with a specific emotion. We blurred the faces of the movers in those clips to avoid any indication for their emotional state through facial expressions, and in our next experiment [42], we asked 62 participants to observe these clips and indicate which emotion they perceive from the movement they observed. Note that the Laban experts who moved these combinations of motor components were not asked to express emotions, but just to move movements which include various combinations of motor components. They were not told what was the purpose of the study, nor were they told about the results of our previous experiments or the hypothesized association between the motor components they were asked to move and emotions. The results showed that all emotions were recognized from these clips above chance level. Moreover, to ensure that all the participants had normal capability for emotion recognition from movements, we also asked them to recognize the emotions expressed in Atkinson's validated clips [4]. Comparing emotion recognition from our 'combinations clips' to that from Atkinson's emotional expression clips, we found that although the movers in our clips were not asked to express emotions, happiness and sadness were recognized from our 'combinations clips' as good as from Atkinson's clips, in which the actors were specifically asked to express emotions.

In sum, in a series of studies we found the motor components that are associated with bodily expressions of each of the four basic emotions: anger, fear, sadness and happiness, and whose existence in a movement causes that movement to be perceived as expressing the associated emotion.

18.6 Using Laban Motor Components for Automatic Emotion Perception from Movements

Once we identified the Laban motor components which are associated with each emotion, we asked each of 6 Certified Laban Movement Analysts (CMAs) to move about 80 different combinations of these components. The resulted movement sequences created a data set of about 550 movement sequences that were filmed each, using both a Kinect camera and an ordinary RGB video camera. In addition, we directed 2 non-CMA people to move some of these combinations and they created another 30 movement sequences. We then used machine learning techniques to teach the computer to automatically identify these motor components in the Kinect 3D skeletal data [6, 7]. To the best of our knowledge, this is the only study where machine learning was used to learn to identify Laban motor components as output, using a data set of videos of Laban components. In most studies that used Laban components, they were used for teaching the machine to identify emotions in bodily emotional expressions, and the Laban components were used as an inspiration for creating the features used during the machine learning process for recognizing the expressed emotions, but were not identified by themselves (e.g., [3, 46, 61]). In these studies, the same Laban motor components were often “translated” in different studies into different features. For example, Larboulette [37] describes two different ways to compute the Weight Effort: As the sum of the kinetic energy of joints composing the body part that moves (based on [29]), and as the deceleration over a time interval (based on [35]). Ajili [2] used the mean, standard deviation and range of acceleration to present the Weight Effort, Ono [46] used the vertical position of the head to define Weight Effort, while Aristudo [3] identified Weight Effort by studying how the deceleration of motion of the root joint varies over time. According to Aristudo, having peaks in decelerations indicates a movement with Strong Weight, while having no peaks indicates a movement with Light Weight, and he emphasized that Weight is velocity independent. This variety of different features, all supposed to represent the same LMA motor component, suggest that probably many of them if not all of them don’t really capture the actual and complete essence of the Weight Effort.

We believe that in order to teach the computer to identify in any movement the Laban motor components associated with each specific emotion, more studies will be needed to determine which feature(s) represent the best and most accurately each of the Laban components. Moreover, new methods of deep learning which are not using pre-determined features, can also be used to teach the computer to identify

Laban motor components. Once computers will be able to accurately detect the presence of those Laban motor components in any given movement, they will be able to infer which emotion is expressed and its intensity, based on the specific combination of the detected Laban components and the amount or intensity of their presence in the movement, as well as the knowledge of which motor components are associated with which emotion. Moreover, as we saw in our studies, some motor components were detected in movements that expressed different emotions: Sudden Time Effort was associated with both happiness and anger. As mentioned both by Noroozi [45] and Ebdali [19] emotions can be modeled as discrete/categorical, as dimensional or as componential. However, it is very difficult to model dimensional or componential emotions. The appearance of Sudden Time Effort in the expressions of both happiness and anger, might indicate that this motor component is not associated with a discrete emotion, but rather with a certain “emotional trait” or a type of an emotional dimension. Both happiness and anger are considered to be “approach emotions” and they both are characterized by high arousal (although they differ in their valence), and it is possible that Sudden Time Effort is associated with this emotional quality which is mutual to these two emotions. Thus, Laban motor components might be especially suitable for modeling emotional expressions, when the emotions are described as dimensional. In addition, combinations of motor components, each of which is associated with a different emotion, can be used to model and/or to identify componential or mixed emotions.

18.7 Conclusion

While most studies of automatic emotion-recognition from body movement used features that described various aspects of movements, often inspired by Laban Movement Analysis, as input, to identify/recognize the emotion expressed by the movement as output, we suggest a different approach: To use machine learning techniques for automatic recognition of Laban motor components as output, while establishing the relationships between certain Laban motor components and specific emotions using behavioral studies. In this chapter we described our behavioral studies which associated certain Laban motor components with the 4 basic emotions: anger, fear, happiness and sadness, and our first study of using machine learning to detect those Laban motor components in people’s movements. More behavioral studies are needed to strengthen and refine our findings and to determine the Laban motor components associated with additional emotions, and more machine learning studies are needed to improve and refine automatic recognition of Laban motor components. Nevertheless, we believe that in the long term, this strategy of research will create a more efficient and more accurate automatic emotion recognition from people expressing their emotions in the wild. In addition, it will enable to better perceive the intensity of the emotion as well as better identify compound emotions or mixtures of emotions.

We would like to conclude by mentioning three recent studies which support our results, indicating the strength and potential of our approach: The first is a study by Ajili et al. [2], in which they developed a machine learning technique to extract from videos several LMA motor components. They then used this technique to extract those LMA components from a data set of emotional expressions as well as asked viewers to rate the type of emotion expressed in each clip of that data set, and calculated the correlation between the LMA components and the rated emotion. Some of the strong correlations that they found indicate the same associations between certain emotions and specific Laban components as those that we found: Similar to our results, they too found anger to be associated with Strong and Sudden movements, happiness to be associated with Spreading the body, and sad to be associated with shrinking (=sinking in Laban terms) and small Kinesphere [2]. The second study which supports our results is a brain imaging study performed at de Gelder's laboratory, in which they found associations in the brain between fear and Condensing the body [49]. Condense is one of the motor components that we found as associated with fear. These studies confirmed the existence of some of the associations that we found between certain emotions and specific LMA motor components, supporting the future high potential of our proposed approach for developing automatic recognition of emotions. Lastly, in a recent study, we introduced a new automatic emotion recognition method: the MANet (Movement Analysis Network) which extracts relevant features of LMA components in one branch and relevant features of emotions in another branch. It then integrates the emotion features with the LMA features to generate its final automatic emotion recognition. Testing this method with videos of bodily expressions of sadness and happiness produced better results compared to using state-of-the-art methods described in other recent studies [64].

References

1. Ahmed, F., Bari, A.H., Gavrilova, M.L.: Emotion recognition from body movement. *IEEE Access* **8**, 11761–11781 (2019)
2. Ajili, I., Mallem, M., Didier, J.Y.: Human motions and emotions recognition inspired by LMA qualities. *Vis. Comput.* **35**(10), 1411–1426 (2019)
3. Aristidou, A., Charalambous, P., Chrysanthou, Y.: Emotion analysis and classification: understanding the performers' emotions using the LMA entities. In: Computer Graphics Forum, vol. 34, pp. 262–276. Wiley Online Library (2015)
4. Atkinson, A.P., Dittrich, W.H., Gemmell, A.J., Young, A.W.: Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception* **33**(6), 717–746 (2004)
5. Barliya, A., Omlor, L., Giese, M.A., Berthoz, A., Flash, T.: Expression of emotion in the kinematics of locomotion. *Exp. Brain Res.* **225**, 159–176 (2013)
6. Bernstein, R., Shafir, T., Tsachor, R., Studd, K., Schuster, A.: Laban movement analysis using kinect. *Int. J. Comput. Inf. Eng.* **9**(6), 1567–1571 (2015)
7. Bernstein, R., Shafir, T., Tsachor, R., Studd, K., Schuster, A.: Multitask learning for Laban movement analysis. In: Proceedings of the 2nd International Workshop on Movement and Computing, pp. 37–44. ACM, Vancouver, BC (2015)

8. Camurri, A., Lagerlöf, I., Volpe, G.: Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques. *International J. Hum. Comput. Stud.* **59**(1–2), 213–225 (2003)
9. Coles, N.A., Larsen, J.T., Lench, H.C.: A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychol. Bull.* **145**(6), 610 (2019)
10. Crane, E.A., Gross, M.M.: Effort-shape characteristics of emotion-related body movement. *J. Nonverbal Behav.* **37**, 91–105 (2013)
11. Cruz-Garza, J.G., Hernandez, Z.R., Nepaul, S., Bradley, K.K., Contreras-Vidal, J.L.: Neural decoding of expressive human movement from scalp electroencephalography (EEG). *Front. Hum. Neurosci.* **8**, 188 (2014)
12. Cui, H., Maguire, C., LaViers, A.: Laban-inspired task-constrained variable motion generation on expressive aerial robots. *Robotics* **8**(2), 24 (2019)
13. Dael, N., Mortillaro, M., Scherer, K.R.: Emotion expression in body action and posture. *Emotion* **12**(5), 1085 (2012)
14. Damasio, A.R.: *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Houghton Mifflin Harcourt (1999)
15. Damasio, A., Carvalho, G.B.: The nature of feelings: evolutionary and neurobiological origins. *Nature Rev. Neurosci.* **14**(2), 143–152 (2013)
16. Damasio, A.R., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L., Parvizi, J., Hichwa, R.D.: Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nat. Neurosci.* **3**(10), 1049–1056 (2000)
17. Darwin, C., Prodgger, P.: *The Expression of the Emotions in Man and Animals*. Oxford University Press, USA (1998)
18. De Meijer, M.: The contribution of general features of body movement to the attribution of emotions. *J. Nonverbal Behav.* **13**, 247–268 (1989)
19. Ebdali Takalloo, L., Li, K.F., Takano, K.: An overview of emotion recognition from body movement. In: *Complex, Intelligent and Software Intensive Systems: Proceedings of the 16th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS-2022)*, pp. 105–117. Springer (2022)
20. Fagan, R., Conitz, J., Kunibe, E.: Observing behavioral qualities. *Int. J. Comp. Psychol.* **10**(4) (1997)
21. Fanselow, M.S.: Emotion, motivation and function. *Curr. Opin. Behav. Sci.* **19**, 105–109 (2018)
22. Foroud, A., Pellis, S.M.: The development of “roughness” in the play fighting of rats: A Laban movement analysis perspective. *Dev. Psychobiol. J. Int. Soc. Dev. Psychobiol.* **42**(1), 35–43 (2003)
23. Foroud, A., Whishaw, I.Q.: Changes in the kinematic structure and non-kinematic features of movements during skilled reaching after stroke: A laban movement analysis in two case studies. *J. Neurosci. Methods* **158**(1), 137–149 (2006)
24. Foroud, A., Whishaw, I.Q.: The consummatory origins of visually guided reaching in human infants: a dynamic integration of whole-body and upper-limb movements. *Behav. Brain Res.* **231**(2), 343–355 (2012)
25. Gallesse, V., Sinigaglia, C.: What is so special about embodied simulation? *Trends Cogn. Sci.* **15**(11), 512–519 (2011)
26. Gendolla, G.H.: On the impact of mood on behavior: An integrative theory and a review. *Rev. Gen. Psychol.* **4**(4), 378–408 (2000)
27. Gross, M.M., Crane, E.A., Fredrickson, B.L.: Methodology for assessing bodily expression of emotion. *J. Nonverbal Behav.* **34**, 223–248 (2010)
28. Gross, M.M., Crane, E.A., Fredrickson, B.L.: Effort-shape and kinematic assessment of bodily expression of emotion during gait. *Hum. Movement Sci.* **31**(1), 202–221 (2012)
29. Hachimura, K., Takashina, K., Yoshimura, M.: Analysis and evaluation of dancing movement based on LMA. In: *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication*, 2005, pp. 294–299. IEEE (2005)

30. Heberlein, A.S., Atkinson, A.P.: Neuroscientific evidence for simulation and shared substrates in emotion recognition: beyond faces. *Emotion Rev.* **1**(2), 162–177 (2009)
31. Huis In 't Veld, E.M., van Boxtel, G.J., de Gelder, B.: The body action coding system II: muscle activations during the perception and expression of emotion. *Front. Behav. Neurosci.* **8**, 330 (2014)
32. Huis In 't Veld, E.M., Van Boxtel, G.J., de Gelder, B.: The body action coding system I: Muscle activations during the perception and expression of emotion. *Soc. Neurosci.* **9**(3), 249–264 (2014)
33. Inthiam, J., Hayashi, E., Jitviriya, W., Mowshowitz, A.: Development of an emotional expression platform based on lma-shape and interactive evolution computation. In: 2018 4th International Conference on Control, Automation and Robotics (ICCAR), pp. 11–16. IEEE (2018)
34. Izard, C.E.: The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Rev.* **2**(4), 363–370 (2010)
35. Kapadia, M., Chiang, I.K., Thomas, T., Badler, N.I., Kider Jr, J.T.: Efficient motion retrieval in large motion databases. In: Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, pp. 19–28 (2013)
36. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: A survey. *IEEE Trans. Affect. Comput.* **4**(1), 15–33 (2012)
37. Larboulette, C., Gibet, S.: A review of computable expressive descriptors of human motion. In: Proceedings of the 2nd International Workshop on Movement and Computing, pp. 21–28 (2015)
38. Lerner, J.S., Li, Y., Valdesolo, P., Kassam, K.S.: Emotion and decision making. *Annu. Rev. Psychol.* **66**, 799–823 (2015)
39. Levy, J.A., Duke, M.P.: The use of Laban movement analysis in the study of personality, emotional state and movement style: An exploratory investigation of the veridicality of “body language”. *Individ. Differences Res.* **1**(1) (2003)
40. Lourens, T., Van Berkel, R., Barakova, E.: Communicating emotions and mental states to robots in a real time parallel framework using Laban movement analysis. *Robot. Auton. Syst.* **58**(12), 1256–1265 (2010)
41. Masuda, M., Kato, S.: Motion rendering system for emotion expression of human form robots based on laban movement analysis. In: 19Th International Symposium in Robot and Human Interactive Communication, pp. 324–329. IEEE (2010)
42. Melzer, A., Shafir, T., Tsachor, R.P.: How do we recognize emotion from movement? Specific motor components contribute to the recognition of each emotion. *Front. Psychol.* **10**, 1389 (2019)
43. Montepare, J., Koff, E., Zaitchik, D., Albert, M.: The use of body movements and gestures as cues to emotions in younger and older adults. *J. Nonverbal Behav.* **23**, 133–152 (1999)
44. Niedenthal, P.M., Mermilliod, M., Maringer, M., Hess, U.: The simulation of smiles (SIMS) model: Embodied simulation and the meaning of facial expression. *Behav. Brain Sci.* **33**(6), 417–433 (2010)
45. Noroozi, F., Corneanu, C.A., Kamińska, D., Sapiński, T., Escalera, S., Anbarjafari, G.: Survey on emotional body gesture recognition. *IEEE Trans. Affect. Comput.* **12**(2), 505–523 (2018)
46. Ono, Y., Aoyagi, S., Yamazaki, Y., Yamamoto, M., Nagata, N.: Emotion estimation using body expression types based on LMA and sensitivity analysis. In: 2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), pp. 348–353. IEEE (2019)
47. Parteli, L.: Aesthetic listening: contributions of dance/movement therapy to the psychic understanding of motor stereotypes and distortions in autism and psychosis in childhood and adolescence. *Arts Psychother.* **22**(3), 241–247 (1995)
48. Pollick, F.E., Paterson, H.M., Bruderlin, A., Sanford, A.J.: Perceiving affect from arm movement. *Cognition* **82**(2), B51–B61 (2001)
49. Poyo Solanas, M., Vaessen, M., de Gelder, B.: Computation-based feature representation of body expressions in the human brain. *Cerebral Cortex* **30**(12), 6376–6390 (2020)

50. Price, T.F., Harmon-Jones, E.: Embodied emotion: The influence of manipulated facial and bodily states on emotive responses. *Wiley Interdiscip. Rev. Cogn. Sci.* **6**(6), 461–473 (2015)
51. Rett, J., Dias, J., Ahuactzin, J.M.: Laban movement analysis using a bayesian model and perspective projections. *Brain Vis. AI* **4**(6), 978–953 (2008)
52. Roether, C.L., Omlor, L., Christensen, A., Giese, M.A.: Critical features for the perception of emotion from gait. *J. Vis.* **9**(6), 15–15 (2009)
53. Ross, P., Atkinson, A.P.: Expanding simulation models of emotional understanding: the case for different modalities, body-state simulation prominence, and developmental trajectories. *Front. Psychol.* **11**, 309 (2020)
54. Sawada, M., Suda, K., Ishii, M.: Expression of emotions in dance: Relation between arm movement characteristics and emotion. *Percept. Motor Skills* **97**(3), 697–708 (2003)
55. Senecal, S., Cuel, L., Aristidou, A., Magenat-Thalmann, N.: Continuous body emotion recognition system during theater performances. *Comput. Anim. Virtual Worlds* **27**(3–4), 311–320 (2016)
56. Shafir, T.: Movement-based strategies for emotion regulation. In: *Handbook on Emotion Regulation: Processes, Cognitive Effects and Social Consequences*, pp. 231–249 (2015)
57. Shafir, T., Taylor, S.F., Atkinson, A.P., Langenecker, S.A., Zubieta, J.K.: Emotion regulation through execution, observation, and imagery of emotional movements. *Brain Cogn.* **82**(2), 219–227 (2013)
58. Shafir, T., Tsachor, R.P., Welch, K.B.: Emotion regulation through movement: Unique sets of movement characteristics are associated with and enhance basic emotions. *Front. Psychol.* **6**, 2030 (2016)
59. Tsachor, R.P., Shafir, T.: How shall I count the ways? A method for quantifying the qualitative aspects of unscripted movement with laban movement analysis. *Front. Psychol.* **10**, 572 (2019)
60. Wallbott, H.G.: Bodily expression of emotion. *Eur. J. Soc. Psychol.* **28**(6), 879–896 (1998)
61. Wang, S., Li, J., Cao, T., Wang, H., Tu, P., Li, Y.: Dance emotion recognition based on laban motion analysis using convolutional neural network and long short-term memory. *IEEE Access* **8**, 124928–124938 (2020)
62. Witkower, Z., Tracy, J.L.: Bodily communication of emotion: Evidence for extrafacial behavioral expressions and available coding systems. *Emotion Rev.* **11**(2), 184–193 (2019)
63. Wood, A., Rychlowska, M., Korb, S., Niedenthal, P.: Fashioning the face: sensorimotor simulation contributes to facial expression recognition. *Trends in Cogn. Sci.* **20**(3), 227–240 (2016)
64. Wu, C., Davaasuren, D., Shafir, T., Tsachor, R., Wang, J.Z.: Bodily expressed emotion understanding through integrating Laban movement analysis. *Patterns* **4**(10) (2023). <https://doi.org/10.1016/j.patter.2023.100816>
65. Zacharatos, H., Gatzoulis, C., Chrysanthou, Y., Aristidou, A.: Emotion recognition for exergames using laban movement analysis. In: *Proceedings of Motion on Games*, pp. 61–66. Association for Computing Machinery (2013)
66. Zhao, L., Badler, N.I.: Acquiring and validating motion qualities from live limb gestures. *Graph. Models* **67**(1), 1–16 (2005)

Chapter 19

Demographic Differences and Biases in Affect Evoked by Visual Features



Baris Kandemir, Hanjoo Kim, Michelle G. Newman, Reginald B. Adams, Jr., Jia Li, and James Z. Wang

Abstract Visual stimuli influence our affective state and reactions, subsequently shaping our preferences. These interactions fall within the domain of affective computing research. Furthermore, affective state and emotions could be influenced by the demographic differences in affective and aesthetic reactions to visual stimuli. This study aims to examine the extent to which demographic factors, such as gender and culture, moderate the associations between prevalent aesthetic features and a range of emotions, represented by valence, arousal, and dominance dimensions. Additionally, this research delves into the identification of latent demographic groups within our participant pool to gain a more nuanced understanding of the interplay among various demographic factors. To accomplish this, more than 40,000 images were collected from web albums and subsequently rated through crowdsourcing based on the emotional and aesthetic response they evoke. The findings of this study reveal that certain colors and aspect ratios elicited distinct valence and dominance reactions depending on the gender of the participant. Furthermore, disparities between Western and Eastern cultures emerged in the relationship between mean brightness and arousal. Latent group analyses identified four

B. Kandemir (✉)

NVIDIA Corporation, Santa Clara, CA, USA

H. Kim

Department of Psychiatry, Michigan Medicine, University of Michigan, Ann Arbor, MI, USA

e-mail: hanjokim@med.umich.edu

M. G. Newman · R. B. Adams, Jr.

Department of Psychology, The Pennsylvania State University, University Park, PA, USA

e-mail: mgn1@psu.edu; rba10@psu.edu

J. Li

Department of Statistics, The Pennsylvania State University, University Park, PA, USA

e-mail: jiali@stat.psu.edu

J. Z. Wang

Data Science and Artificial Intelligence Area, and Human-Computer Interaction Area, College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA

e-mail: jwang@ist.psu.edu

primary groups that exhibited differential responses to specific visual attributes. This investigation contributes to the growing body of literature on affective computing, offering valuable insights into the role of demographic factors in shaping emotional and aesthetic responses to visual stimuli.

19.1 Introduction

Daily choices, including selecting attire, sharing images with acquaintances, or purchasing artwork, involve our aesthetic engagement with the environment. The intricate mechanisms underpinning aesthetic inclinations, intertwined with our emotional states, have piqued the interest of the scientific community. One's affective state can shape the manner in which individuals sense or perceive their surroundings and respond to environmental changes [17]. Visual signals serve as a critical input for our perceptual system. A specialized branch of psychology is devoted to examining the relationship between visual stimuli and emotional reactions [2]. Correspondingly, the domain of *affective computing*, focusing on the development of machines capable of recognizing, modeling, conveying, and responding to emotional information, has garnered significant attention [1].

The emotion theory posits that specific attributes of a visual stimulus determine the nature of the evoked emotion. Consequently, the appraisal processes underlying our aesthetic predilections are influenced by the interplay between emotional state and perception. This interaction results in varying preferences for visual stimuli [7]. The emotional processing of visual cues may function differently for individuals with diverse backgrounds [19]. In alignment with this distinction, marketing research indicates that gender and culture represent two of the most significant dimensions for consumer segmentation [6, 8]. These preferential disparities are taken into account when presenting products to optimize sales in marketing research. Investigations utilizing survey methodologies have further explored gender differences in preferences for color and other design elements, such as aspect ratio. The findings hold relevance for fields including marketing, brand management, decision-making, human-computer interaction, and neurophysiology [3, 9, 10, 18].

We examined the interplay of *visual stimuli*, *emotions*, and *preferences* from a computational standpoint, integrating the automated extraction of features from visual stimuli and demographic difference studies through the power of crowdsourcing. In contrast to prior research conducted in controlled settings, such as laboratory environments, our primary objective was to identify visual cues from psychological studies that may elicit emotionally distinct responses across genders and cultures, and to validate these cues with a larger sample size in real-world settings. Assessing these findings on a broader scale is crucial to determine their validity. Moreover, we delved deeper into other demographic factors, such as age, ethnicity, education, and income, by scrutinizing latent demographic groups and their responses for a more comprehensive understanding. Our study's significance stems from its investigation

of visual affective features prominently employed in computational affect and aesthetic research. The potential applications of our findings in fields such as marketing, e-commerce, and human-computer interaction highlight the importance of this research. Furthermore, the observed gender differences in emotional responses may suggest varying treatment outcomes in affective experiments.

The *main contributions* of this study can be outlined as follows:

- We examined the influence of specific visual features on emotional responses from diverse genders and cultures using a larger dataset obtained through crowdsourcing.
- We employed natural images to evaluate the generalizability of previous research findings derived from controlled laboratory experiments.
- We demonstrated the existence of distinct latent demographic groups among participants.
- We revealed variations in affective and aesthetic responses to images across these demographic groups.

We observed statistically significant differences across gender and cultural dimensions in the relationships between several extracted visual features and emotional constructs. The subsequent sections of the chapter provide a detailed account of the collected data, the visual features extracted, the analytical methods employed, and the results obtained. The chapter concludes with a discussion of the findings and a summary of the implications.

19.2 Data Collection and Feature Extraction

In order to capture the biases and differences in the affective responses among diverse crowdsourcing populations, we collected visual data including elicited emotions and aesthetic ratings from a participant pool sourced from Amazon Mechanical Turk (AMT). This section delves into the methodology employed for data collection and analysis, building upon the dataset presented by Lu et al. [11]. We address the demographic component that was not thoroughly explored in the original dataset by examining demographic disparities.

19.2.1 Human Subject Study

We collected a total of 49,967 medium-sized images (approximately 500×500) from the photo-sharing platform Flickr, ranking them according to an interestingness measure from the same site. Emotional ratings, in terms of constructs of valence, arousal, and dominance, along with aesthetic likeability ratings were collected from 4148 human subjects. The sample comprised 2236 females and 1912 males, with ages ranging from 18 to 72.

Initially, participants were requested to complete a demographic information form, providing details about their gender, age, ethnicity, nationality, income, and education. Ethnicity categories were recorded using a drop-down menu with seven options:

- American Indian or Alaska Native,
- Asian,
- African American,
- Native Hawaiian or Other Pacific Islander,
- Hispanic or Latino,
- Not Hispanic or Latino,
- Other, which included an additional column in which the user was able to specify a value.

Income data were collected using a drop-down menu with income ranges from \$0 to \$100,000 in \$10,000 increments and two other options, \$100,000–\$149,000 and \$150,000 or more. Educational attainment of each user was documented using a categorized drop-down list. The categories were as follows:

- No schooling completed,
- Nursery school to 8th grade,
- 9th, 10th, or 11th grade,
- 12th grade, no diploma,
- High school graduate—diploma or equivalent,
- Some college credit, but less than 1 year,
- 1 or more years of college, no degree,
- Associate's degree,
- Bachelor's degree,
- Master's degree,
- Professional degree,
- Doctorate degree.

19.2.2 Data Cleaning

As we only checked for the involvement of the participants during the survey, we applied a data-cleaning algorithm to exclude data from uninformative participants. For this purpose, we adopted a state-of-the-art technique based on probabilistic graphical models, as described in Ye et al. [20]. Their method is a probabilistic multigraph approach to consensus modeling. The agreement among individual subjects is modeled through a multigraph, incorporating two latent variables: the subject's reliability, following a Bernoulli distribution, and the extent to which the particular subject's response agrees with other reliable responses, following a Beta distribution. The parameters for these latent distributions are obtained through an expectation maximization (EM) framework.

In total, we collected ratings for 41,255 images from 2063 participants. The gender distribution included 1094 females (coded as 1) and 969 males (coded as 0). The two prevailing ethnic groups were “Not Hispanic or Latino” and “Asian.” Income distribution displayed a skew toward participants earning less than \$10,000. The majority of participants held bachelor’s degrees or had completed at least one year of college without obtaining a degree (i.e., undergraduates). Participant ages ranged from 16 to 72, with a skew towards 25 years old, and the predominant age group was 25–35.

Emotion measures were calculated using a scale ranging from 1 to 9. *Valence* exhibited a mean of 5.7 and a standard deviation of 1.9, while *arousal* demonstrated a mean of 5.2 and a standard deviation of 2.0. *Dominance* presented a mean of 4.9 with a standard deviation of 1.8. Figure 19.1 shows the distribution of demographic elements.

19.2.3 Visual Features

To establish a connection between images and emotional responses, it is essential to characterize images through pertinent visual feature extraction. The selected features are derived from those commonly employed in affective computing and computational aesthetics domains. These features not only encompass color, illuminance, and texture attributes, but also offer insights into the photographic and

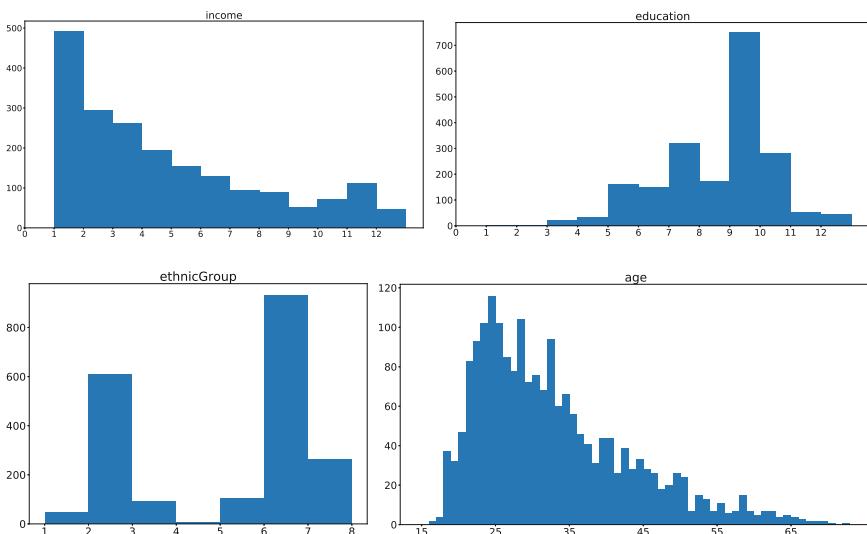


Fig. 19.1 Income, education, ethnicity, and age distributions of AMT participants after data cleaning

Table 19.1 The affective and aesthetic visual features utilized in gender difference analysis

Category	Name	Description
Color	Use of light	Mean brightness
	Pleasure, Arousal, Dominance	Valence, arousal, and dominance values calculated according to regression model
	Colorfulness	Earth Mover Distance (EMD) between the hue histogram of the image and uniform distribution
	Color names	Percentage of black, brown, blue, gray, green, orange, pink, purple, red, white, and yellow
	Itten contrasts	Contrast of brightness and saturation according to Itten's color space
	Harmony	Harmony of colors according to Itten's color accordance
Texture	Wavelet texture	Daubechies' wavelet coefficients for HSV color space for levels 1–3 and their summation for each channel
	Gray-level co-occurrence matrix	Features of contrast, correlation, energy, and homogeneity based on the co-occurrence matrix
Composition	Low depth of field	The ratio of the amount of detail in the rectangle encompassing the image center to the amount of detail in the center section of the image in the whole image for each channel
	Rule of thirds	Average hue, saturation, and value for the center rectangle
Shape	Roundness	How much the image is dominated by round structures
	Angularity	How much the image is dominated by concave structures
	Complexity	Number of segments in the image
Dynamics	Percentage of edge pixels	The percentage of edge pixels representing how dynamic the picture is
Other	Aspect ratio	The ratio of the width to the height of an image

emotional aspects of an image [5, 11, 12, 16]. A comprehensive list of the extracted features can be found in Table 19.1.

19.3 Data Analysis

Our analysis comprised of two distinct approaches. Firstly, we wanted to control for the single effect of gender and culture. In accordance with the norms of psychological analysis in gender studies, which typically involve equal numbers of male and female participants, we opted for a subset of the dataset [13]. However, addressing the imbalance in culture through subset selection proved unfeasible, as it would significantly reduce the sample size. Secondly, we aimed to identify latent groups inherent in our participant base, necessitating an evaluation of the entire dataset *as is*.

19.3.1 Gender and Culture Analysis

This analysis examines the interplay between visual features, gender, and culture in relation to affective and aesthetic responses. While gender is traditionally treated as a binary construct (1: female/high gender, 0: male/low gender in our study), culture is harder to define. In our study, we attempt to define culture as a binary construct by dividing the participant base into Eastern and Western cultural groups based on nationality. In our analysis, we categorized nationalities westwards from Eurasia, inclusive of Eurasia itself, as Western culture, and those to the east of Eurasia as Eastern culture.

19.3.1.1 Multi-Stage Regression Models

Imbalance as a Challenge A primary challenge in our analysis was addressing the imbalanced ratings for each image. For instance, 97.55% of images had an unequal distribution of cultural affiliation, while only 2.45% (1538) had an equal number of Eastern and Western ratings. We analyzed the effects of gender and culture separately. Ensuring balanced ratings from different genders and cultures is crucial for the robustness of statistical analysis.

To address the gender imbalance, we selected images with a perfect gender balance in terms of ratings. Hence, we identified 3078 images rated by 955 female and 867 male participants, yielding a total of 12,600 ratings. Implementing the same approach for culture labeling proved unsatisfactory, as the number of image cases and users declined drastically. Therefore, we developed an alternative measure to address the imbalance. Consequently, gender and culture analyses will be reported separately. For gender analysis, we employed the perfectly balanced rating subset data, while cultural differences were analyzed by devising variables representing the imbalance of culture and gender in the entire dataset.

Before extracting the gender-balanced data, in order to test whether this unequal structure had significant effects on emotional responses, we created gender imbalance and culture imbalance variables for an image, defined as $\text{gender}_{im} = \frac{n_m - n_f}{n_m + n_f}$, where n_m (or n_f) is the number of male (or female) ratings for that image, and $\text{culture}_{im} = \frac{n_w - n_e}{n_w + n_e}$, where n_w (or n_e) is the number of Western (or Eastern) ratings. To assess the impact of imbalances, we tested linear regression models using emotional ratings as response variables and imbalance values as predictor values. Serial linear regression analyses revealed a significant impact of gender imbalance on dimensional ratings of emotion, while culture imbalance demonstrated no significant impact. Hence, we controlled for imbalance in our culture analysis. However, for gender analysis, gender-balanced data was utilized due to the extreme nature of the imbalance. In all subsequent discussions, significance is asserted at the 0.05 level (two-tailed).

Hierarchical Regression For a more accurate analysis, we controlled for culture imbalance by entering them at stage one as the independent variable of each regression model. Along with culture imbalance variables, other demographic variables such as age, education level, and annual income were entered into stage one as control variables. This approach was employed to isolate the independent effects of gender and culture, as they might interact with other demographic variables. In the later stages, multiple models were fitted with respect to independent and interaction effects. Specifically, in the second stage, the model included a particular visual feature, gender, and culture as predictor variables. In the third stage, pairs of interactions (visual feature \times gender, gender \times culture, visual feature \times culture) between the predictors were entered. In the fourth stage, interaction across all three predictors (visual feature \times gender \times culture) was entered as a predictor. In this model, emotional constructs such as valence, arousal and dominance were designated as dependent variables. The regression hierarchy can be summarized as

$$\text{dependent var.} = \beta_{\text{control1}} \times \text{gender imbalance} + \beta_{\text{control2}} \times \text{culture imbalance} \\ + \beta_{\text{control3}} \times \text{control var.}, \quad (19.1)$$

$$\text{dependent var.} = \beta_{\text{control}} \times \text{control var} + \beta_{\text{main}} \times \text{visual feat.} + \beta_{\text{gender}} \times \text{gender} \\ (\text{or } + \beta_{\text{culture}} \times \text{culture}), \quad (19.2)$$

$$\text{dependent var.} = \beta_{\text{control}} \times \text{control var.} + \beta_{\text{gender}} \times \text{gender} (\text{or } + \beta_{\text{culture}} \times \text{culture}) \\ + \beta_{\text{main}} \times \text{visual feat.} + \beta_{\text{interactions}} \times \text{interactions}. \quad (19.3)$$

with variables from Stage 1 carried over to Stage 2 as control variables.

In the following part, the visual feature name, e.g., *blue*, represents the model in either (19.2) or (19.3), where the dependent variable was one of the emotional constructs, and the relevant visual feature served as an independent variable alongside its interaction with gender or culture variable. All models were significant in predicting each emotional construct, thereby providing evidence for the association between colors, brightness, aspect ratio, and emotions.

19.3.1.2 Results of Multi-Stage Regressions

Main and Interaction Effect of Gender Across all emotional constructs, the primary independent effect of gender was not significant across all models except for arousal and likeness. Female participants generally experienced lower arousal and greater attraction to images.

The relationship between visual components and gender demonstrated positive significant interaction effects across various types of models. The interaction effect coefficients between blue and gender, as well as aspect ratio and gender, were significantly positive on valence. This suggests that as images become bluer or more horizontally elongated, female participants are more likely to experience

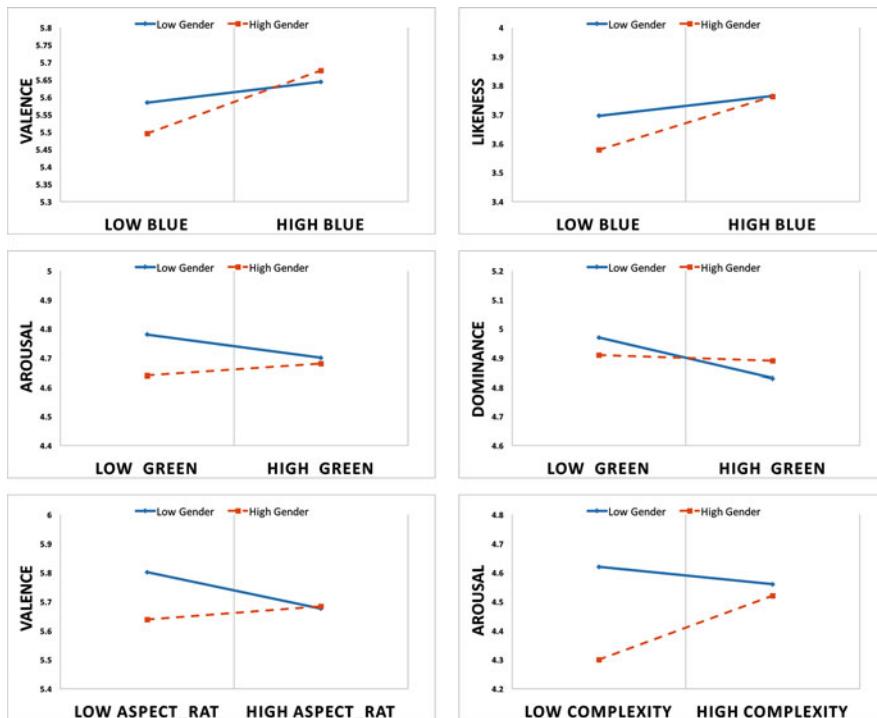


Fig. 19.2 The interaction of visual features with gender on various emotional constructs and likeness, with high gender representing females and low gender representing males

increased positive valence. Green exhibited significant interaction with gender in terms of arousal and dominance. According to the models, female participants' reactions to greener images were characterized by more pronounced changes in energy and emotional captivation. A noteworthy interaction was observed between complexity and gender in relation to arousal. As images become more complex, female participants' emotions become more energetic in a more drastic manner. Figure 19.2 depicts the significant interaction effects discovered. The positive interaction effect of aspect ratio implies that more landscape-oriented images evoke more positive emotions in female participants compared to their male counterparts.

Main and Interaction Effect of Culture Compared to gender, culture presented more consistent main effects across all models. The main effect coefficient of culture was positive on both valence and arousal. This indicates that participants originating from Eastern cultures reported significantly higher valence and arousal than their Western counterparts. Conversely, culture indicated a significant negative effect on dominance, indicating that Western participants experienced a greater sense of control when viewing a picture.

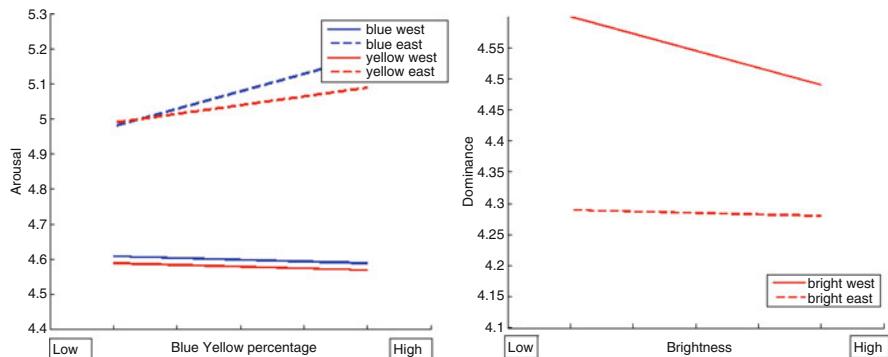


Fig. 19.3 The interaction of blue, yellow, and culture on arousal, and the interaction of brightness and culture on dominance

Interaction effect coefficients between picture components and culture were predominantly positive, with the exception of red. Among visual components, blue, green, yellow, brightness, and aspect-ratio exhibited significant positive interactions with culture when arousal served as the dependent variable (Fig. 19.3). However, red and culture yielded a significant negative effect on arousal, revealing a difference in reactions to red-dominated stimuli across cultures. Alternatively, when dominance acted as the dependent variable, the interaction between brightness and culture also demonstrated a significant positive effect, as illustrated in Fig. 19.3. A three-way interaction between visual components, gender, and culture was not significant in this study.

19.3.2 A Comprehensive Approach to Demographic Disparities

Our investigation of the full dataset, divided based on a single demographic variable, failed to produce any robust correlation results concerning disparities, as other variables were not controlled for. The world is not linear, and seeking linear relationships through regression analysis may not be appropriate for this type of research. An analysis that simultaneously considers multiple demographic variables to characterize different demographic groups is necessary.

This calls for a method that categorizes similar individuals into a single classification, aligning with the fundamental concept of clustering. A demographic cluster can be defined in the same way as we capture the topic of a document, using a set of keywords such as “young low-income graduates.” This similarity can be identified by employing a topic modeling approach to discover latent abstract groups within the AMT population sample that we obtained.

19.3.2.1 Latent Demographics Group

Latent Dirichlet Allocation (LDA), one of the topic modeling methods, is a generative probabilistic model of a corpus. In this technique, documents are represented as random mixtures over latent topics, which are portrayed by a distribution over a word [4]. LDA was employed in [16] to characterize the favorite images of individuals with certain personality traits. Similarly, our objective was to characterize abstract groups from the samples. Thus, we applied LDA to our participant pool. With the exception of age, the demographic information for each participant (i.e., ethnicity, education, income) was already quantized. The quantized demographic values were treated as word frequencies, drawing from the document analysis concept. LDA analysis in the participant pool identified four prominent latent groups. The description column in Table 19.2 provides details on these groups. Figure 19.4 displays the distribution of different demographic categories for various latent groups.

Upon identifying the latent demographics groups within the sample, we analyzed participants' responses to the images, aiming to discern correlations between visual features that could indicate affective biases and differences. As the results presented above indicate, selecting one feature and analyzing the correlations between emotional constructs did not yield robust outcomes. This may be attributed to the simultaneous presence of multiple features within an image, which together elicit a response. Consequently, we had to further clump together the images that were rated by each group according to the images' visual similarities in order to pin specific visual features to particular emotional responses.

Following a methodology akin to [16], we utilized LDA to separate the images into latent groups for each demographic group. The visual feature values underwent quantization, and the quantized data were subsequently divided into topics characterized by feature names through LDA. For instance, a topic might be *blue, green, orange, gray, brown, red, yellow, percentage of edge pixels, purple, or use of light*, with the sequence of feature names signifying the contribution of each specific feature to that topic. One example topic is demonstrated in Fig. 19.5. We observed that color features played a dominant role in forming the visual topics, primarily due to the dataset's composition of predominantly color images. We conducted the same topic analysis without incorporating color features to determine the influence of other features in representing visual topics. Correlation relationships between each emotional and aesthetic construct and each individual feature among the topics were analyzed, and the Spearman correlation coefficients (ρ) were calculated with associated statistical significance values (p).

19.3.2.2 Differences Across Latent Groups

Table 19.2 presents the significant correlations identified between primary visual topics and the constructs of interest. While certain commonalities exist across the groups, the disparities in these correlations merit attention. Group 0 deviates from

Table 19.2 The demographic properties of groups and significant correlations between constructs and visual features. mahogany color signifies negative correlation and cobalt color shows positive correlation at a significance level of $p < 0.05$

Group	Description	Findings			
Group 0	Individuals older than 35. Mostly “not Hispanic, not Latino.” Mainly master’s degree Annual income < \$90,000. females (250) > males (140).	Valence Use of light, green, pink, white brown, orange, complexity, angularity Edge pixel measure,	Arousal Green, gray complexity, yellow, red, orange, brown	Dominance Angularity, gray, purple white, orange, red	Likability Complexity, purple, orange green, yellow
Group 1	Individuals between 30–35 years old (1)“not Hispanic, not Latino.”, (2)Asians. Bachelor’s degree. Annual income \$100,000–\$149,999. # Female = # male.	brown, orange, red complexity, blue rule of thirds, brown, gray, orange, white, purple, complexity	Green, edge pixel measure, Red, complexity, angularity, rule of thirds green, texture homogeneity-saturation	Purple	
Group 2	Individuals between 19–32 years old. “not Hispanic, not Latino.” Bachelor’s degree, or 1+ years of college, no degree. Annual income < \$70,000. # Female = # male.	Brown, pink, orange blue, yellow, purple green, white, complexity	Brown, gray, colorfulness yellow, red, complexity	Complexity, angularity, gray, purple, use of light, pink, blue std. of brightness blue, edge pixel measure texture homogeneity-saturation	Purple, orange
Group 3	Individuals between 25–30 years old. Asian. (1) Bachelor’s, or (2) master’s degrees. Annual income < \$40,000. # Female << # male.	Green, white angularity, yellow, purple	Level of detail complexity, blue, purple brown, orange, red, purple	Red, textural contrast-saturation complexity, orange angularity, level of detail	

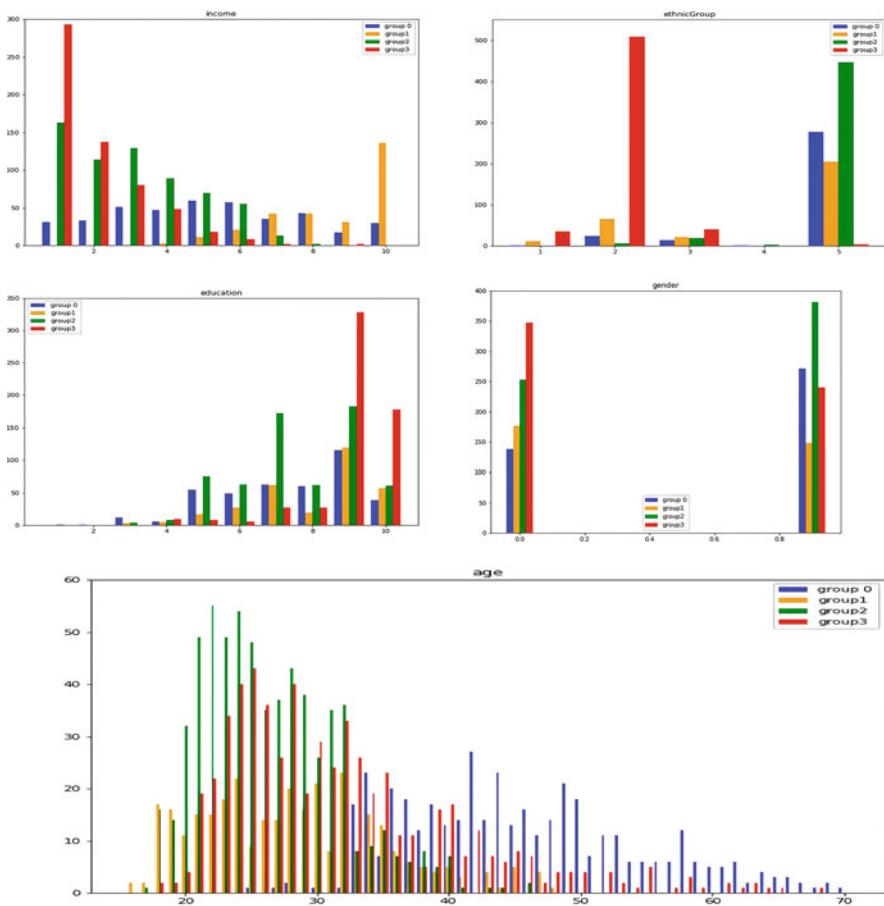


Fig. 19.4 Latent representative demographic groups and their make-up in terms of income, ethnic group, education, gender, and age

other groups by exhibiting a positive correlation between the brown color and valence. Group 2 sets itself apart by revealing positive correlations for both white-valence and green-valence associations. The blue-valence relationship distinguishes Groups 1 and 2, while the purple-valence correlation differentiates Groups 2 and 3.

Group 3 diverged from the other groups with a negative brown-arousal link. Group 1 contrasts with Groups 0 and 2 by demonstrating a positive correlation between gray and arousal. The white-arousal link separates Groups 1 and 3. Furthermore, Groups 0 and 1 displayed divergent relationships between red and dominance, and Group 3 exhibits a distinct relationship between blue and dominance compared to Groups 1 and 2. The light-dominance correlation varies between Groups 0 and 2.

The percentage of edge pixels reveals different correlation relations with dominance among Group 0 and Groups 2 and 3. Additionally, the green-dominance

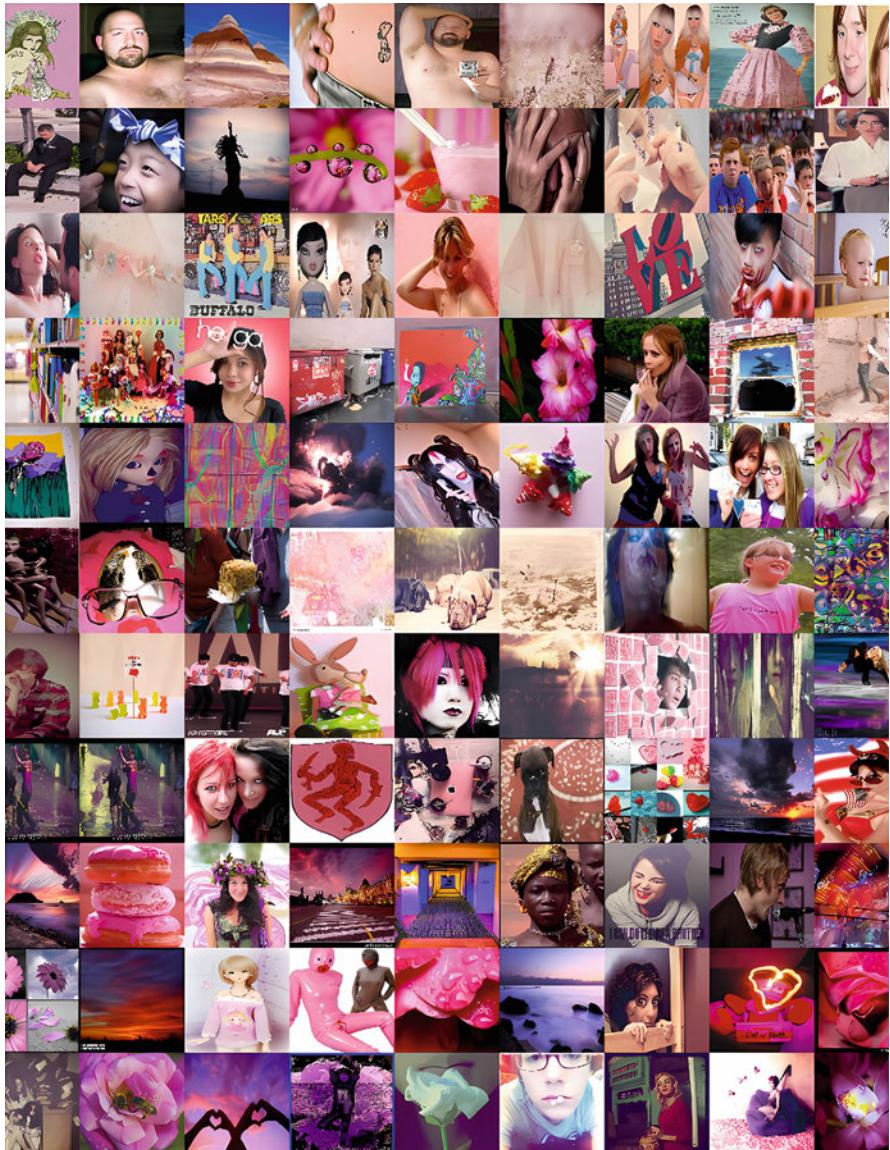


Fig. 19.5 A sample of a topic whose most dominant words were purple, red, and white. Each image that belongs to this topic was ordered via the vectorial norm of the three features' values. Lower-triangle images demonstrate these features better. The images are sourced from Flickr, and all associated copyrights are retained by the original content creators. They are incorporated in the figure for illustrative purposes and to support the conceptual discussions in this chapter. The authors acknowledge and appreciate the work of the original content creators

relation causes Group 1 to differ from Group 3. While Groups 0 and 2 share negative associations between orange and likability, Groups 1 and 3 displayed positive relations. Lastly, the correlation between gray and likability sets Groups 1 and 2 apart.

19.4 Discussion

Our statistical findings from the first part on gender and color percentage interaction suggest that individuals exhibit positive valenced emotions towards blue-dominated images, with a more pronounced effect among females, indicating a gender difference. These results align with the psychological literature, such as [13], which reports gender differences in the perception of blue. Green significantly interacts with gender on arousal and dominance. According to the models, female participants' reaction to greener images involved sharper changes in energy and captivation of the feelings. A noteworthy interaction between complexity and gender on arousal was observed; as images become more complex, female participants' emotions become more energetic in a more pronounced manner. Figure 19.2 depicts the significant interaction effects identified.

The positive interaction effect of aspect ratio implies that more landscape-oriented images elicit more positive emotions among female participants compared to male participants. Furthermore, brightness demonstrates a cultural difference in dominance; Eastern cultured participants feel more in control when exposed to brighter stimuli compared to their Western counterparts. These result may inform the development of cultural ontologies for interface adaptation [15] and facilitate more effective mapping of cultural cues and interface modification suggestions. Such findings may be applied in multimedia, potentially enabling online companies to tailor advertising content based on demographic information. Cultural differences in the relationship between emotional constructs and visual features could prove valuable for localized commercial Website design.

In the second part of our study, the findings uncover intriguing patterns of agreement and divergence across the study's groups. One aspect warranting further discussion is the methodology employed to separate participants into demographic groups. Alternative grouping methods, such as clustering, are also available. However, in our experience, the simplest clustering method, k-means, produced many small clusters that complicated analysis and rendered it intractable. Treating each participant as a document yielded meaningful separation. In our user study, we only checked user involvement time as a quality check and relied on a computational method to eliminate uninformative workers' data. We could have incorporated intermittent attention check questions during the survey. Peer et al. [14] described how AMT workers become accustomed to attention check questions, providing irrelevant data despite passing the attention check. As the performance of our computational data cleaning approach was better than acceptable, we opted against integrating attention check questions into the survey. Another potential concern is

our decision not to compensate for chance in our correlation analyses. As our study involved more than a thousand correlation analyses, the chance may have affected our results. One correction for chance is Bonferroni's correction. This correction is more or less conservative, and any correction may leave out significant results that may be interesting to the audience. Consequently, we present the results as they stand.

19.5 Conclusions

In this study, we aimed to investigate whether theories regarding preference differences hold true in real-world settings. We recruited a diverse group of participants from around the globe through Amazon Mechanical Turk. Hierarchical regression-based statistical analyses revealed significant gender differences in emotional responses to natural images. We identified latent demographic groups and examined image sets rated by each group. These images underwent additional topic modeling to sense the cluster visual features that exist together in our dataset. We then inspected correlation analyses between visual features and emotional-preferential constructs. Our findings revealed differing responses to specific colors across the groups, while observing consensus on responses to structural features such as complexity and simplicity. This work holds the potential to significantly impact the design of multimedia content, advertising, and user interfaces by providing a deeper understanding of preference differences, ultimately enabling more personalized and culturally sensitive experiences for users across diverse demographic backgrounds.

Acknowledgments This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1110970. The work was done when B. Kandemir and H. Kim were with Penn State. J. Z. Wang has also been supported by generous gifts from the Amazon Research Awards program. The research utilized the Extreme Science and Engineering Discovery Environment, which was supported by NSF under Grant No. ACI-1548562.

References

1. Becker, W.J., Tuskey, S.E., Beugré, C.D.: HR affective computing. In: *Handbook of Research on Artificial Intelligence in Human Resource Management*. Edward Elgar Publishing (2022)
2. Bradley, M.M., Sambuco, N., Lang, P.J.: Affective perception: The power is in the picture. In: *Human Perception of Visual Information*, pp. 59–83. Springer (2022)
3. Butkowski, C.P.: Beyond “commercial realism”: Extending Goffman’s gender display framework to networked media contexts. *Commun. Cult. Critique* **14**(1), 89–108 (2021)
4. Chauhan, U., Shah, A.: Topic modeling using latent Dirichlet allocation: A survey. *ACM Comput. Surv.* **54**(7), 1–35 (2021)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: *Proceedings of the European Conference on Computer Vision*, pp. 288–301. Springer (2006)

6. Holbrook, M.B.: Essay on the origins, development and future of the consumption experience as a concept in marketing and consumer research. *Qualitat. Mark. Res. Int. J.* **21**, 421 (2018)
7. Kuppens, P., Champagne, D., Tuerlinckx, F.: The dynamic interplay between appraisal and core affect in daily life. *Front. Psychol.* **3**, 380 (2012)
8. Lieven, T.: Product gender and product evaluation. In: *Brand Gender*, pp. 143–176. Springer (2018)
9. Lima, A.L.d.S., Gresse von Wangenheim, C.: Assessing the visual esthetics of user interfaces: a ten-year systematic mapping. *Int. J. Hum. Comput. Interaction* **38**(2), 144–164 (2022)
10. Lithari, C., Frantzidis, C., Papadelis, C., Vivas, A.B., Klados, M., Kourtidou-Papadeli, C., Pappas, C., Ioannides, A., Bamidis, P.: Are females more responsive to emotional stimuli? A neurophysiological study across arousal and valence dimensions. *Brain Topogr.* **23**(1), 27–40 (2010)
11. Lu, X., Adams Jr, R.B., Li, J., Newman, M.G., Wang, J.Z.: An investigation into three visual characteristics of complex scenes that evoke human emotion. In: *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pp. 440–447. IEEE (2017)
12. Machajdik, J., Hanbury, A.: Affective image classification using features inspired by psychology and art theory. In: *Proceedings of the International Conference on Multimedia*, pp. 83–92. ACM (2010)
13. McInnis, J.H., Shearer, J.K.: Relationship between color choice and selected preferences of the individual. *J. Home Econ.* **56**(3), 181–187 (1964)
14. Peer, E., Brandimarte, L., Samat, S., Acquisti, A.: Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *J. Exp. Soc. Psychol.* **70**, 153–163 (2017)
15. Reinecke, K., Bernstein, A.: Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Trans. Comput. Hum. Interaction* **18**(2), 8 (2011)
16. Segalin, C., Perina, A., Cristani, M., Vinciarelli, A.: The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits. *IEEE Trans. Affect. Comput.* **8**(2), 268–285 (2017)
17. Speer, M.E., Delgado, M.R.: Emotion–cognition interactions in memory and decision making. In: *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, vol. 4, pp. 1–26 (2018)
18. Stumpf, S., Peters, A., Bardzell, S., Burnett, M., Busse, D., Cauchard, J., Churchill, E., et al.: Gender-inclusive HCI research and design: A conceptual review. *Found. Trends® Hum. Comput. Interaction* **13**(1), 1–69 (2020)
19. Wrage, J., Klein, S., Gruesser, S.M., Hermann, D., Flor, H., Mann, K., Braus, D.F., Heinz, A.: Gender differences in the processing of standardized emotional visual stimuli in humans: a functional magnetic resonance imaging study. *Neurosci. Lett.* **348**(1), 41–45 (2003)
20. Ye, J., Li, J., Newman, M., Adams Jr, R.B., Wang, J.Z.: Probabilistic multigraph modeling for improving the quality of crowdsourced affective data. *IEEE Trans. Affect. Comput.* **10**(1), 115–128 (2019)

Part VII

Artistic Style

This part explores the intersection of computational analysis and the understanding of fine art.

In the first chapter, “Deep Network-based Computational Transfer of Artistic Style in Art Analysis,” by transferring the styles of representative artworks onto natural photographs, surrogate artworks are created, enabling deep networks to achieve state-of-the-art performance in image segmentation tasks related to fine art.

The second chapter, “Balance of Unity and Variety in Fine Art Paintings: A Computational Study,” investigates the concept of balance in visual artworks. Inspired by the teachings of Harold Speed, the chapter explores the interplay between unity and variety in achieving a balanced composition.

These chapters offer new perspectives on computational approaches to studying fine art.

Chapter 20

Deep Network-Based Computational Transfer of Artistic Style in Art Analysis



David G. Stork

Abstract Some of the most successful computational approaches to the analysis of natural photographs and videos are based on deep neural networks trained with large corpora of representative photographs. Unfortunately, such trained networks have but modest performance on analogous tasks when they are applied directly to fine art paintings and drawings, including semantic segmentation, captioning, and question answering. The obvious program would be to train (or transfer train) existing network architectures with large corpora of representative fine art images but even the largest art corpora are far too small to yield high accuracy on fine art images. We describe an alternate approach in which separate deep networks computationally transfer the style of representative artworks and movements onto large corpora of natural photographs to thereby create *surrogate artworks*. Such surrogate artworks might include a helicopter rendered in the style of Claude Monet or a portrait of Oprah Winfrey rendered in the style of Caravaggio. Separate application-specific deep networks trained with such surrogate artworks show state-of-the-art performance on image segmentation. The accuracy of these networks implies that the early, low-level feature extractors are most altered to apply to fine art. Such creation of surrogate art images should find wide use throughout computational studies of fine art paintings and drawings.

20.1 Background: Deep Network-Based Analysis of Photographs and Fine Art Paintings and Drawings

The emergence of deep neural networks has revolutionized many branches of artificial intelligence, perhaps none as profoundly as that of our central concern: image analysis [9, 15, 49]. Such networks, generally trained with extremely large datasets, develop highly complex spatial, chromatic, and related models leading to

D. G. Stork (✉)
Stanford University, Stanford, CA, USA
e-mail: dstork@stanford.edu

state-of-the-art accuracy on complex tasks such as pattern classification, semantic image segmentation, image summarization, question answering, and others.

It is of course natural that such networks would be applied to problems in the history and interpretation of fine art paintings and drawings, and indeed on some proscribed categories of art analysis problems, such as craquelure detection, this approach has led to state-of-the-art performance [38]. However, as we shall see below, networks trained on large corpora of natural photographs generally exhibit poor performance on higher, cognitive tasks in the analysis of fine art paintings and drawings, including semantic segmentation. At base, this problem can be ascribed to the wide variety of styles among art images and the lack of adequately large and appropriate training sets.

In this chapter we present and discuss recent results that show how computational style transfer, implemented by deep networks, can improve accuracy on the high-level task of semantic segmentation, and argue that the approach should be used in other computational analyses of art [19, 45].

In Sect. 20.2 we discuss background, specifically eight inter-related classes of problems arising in art analysis that are not widely addressed in mainstream artificial intelligence and deep learning research. We shall see that solving the problem of (relatively) small data sets of art images will help solve most or all of these classes of problems. In Sect. 20.3 we focus more specifically on the problem of small data sets, including one-shot (or few-shots) learning, that is, how both humans and current deep networks learn directly with small data sets. We shall see that at present purely algorithmic approaches are inadequate to solve this problem and that progress will rely on large corpora. In Sect. 20.4 we concentrate on our central topic: how to use style transfer to generate large corpora of *surrogate art images* for training networks for high-level tasks of art analysis. In Sect. 20.5 we review deep network approaches to the task of image segmentation. In Sect. 20.6 we show how segmentation networks transfer trained with surrogate art images lead to segmentations of art that are superior to those from networks trained with large corpora of natural photographs alone. We conclude in Sect. 20.7 with a brief summary of some lessons learned and suggestions for future research directions.

20.2 Problems in Artificial Intelligence Specific to the Analysis of Fine Art Paintings and Drawings

Mainstream artificial intelligence research has focussed on a wide range of image-based application areas, often selected because of the fundamental nature of the problems, or the commercial and societal value of the solutions, or sometimes both. Leading applications include autonomous driving, medical image analysis, surveillance, detecting tampering and deep fakes, remote sensing, Esports and games such as go and Texas hold’em, robot and aerial vehicle guidance, and many others. In the past few years, a small but growing interest has arisen around problems posed by art history and criticism [23, 24, 39, 42, 43].

20.2.1 Classes of Challenges to Computational Art Analysis

Here are nine inter-related problems posed exclusively by (or primarily by) the scholarly analysis of fine art paintings and drawings that are not, generally, central to research in mainstream artificial intelligence: [36, 40, 41]

Style Although there is no broadly agreed upon definition of “style” applicable throughout the grand sweep of fine art, there is an agreement among art scholars that style often relates to choices in color, brushstroke, composition, iconography, and much more. Style is an essential property of many artworks, and allows us to group artworks independent of the particular subject matter or “content.”

Style is often an important aspect of the meaning of an artwork as well. For example, the relatively free, quick brush strokes of the Impressionists, often executed en plein air, revealed their preoccupation with recording their impressions of daily life. The highly realistic style of vanitas still lives from the Dutch Golden Age supported readings of these works stressing that the depicted symbols (skulls, watches, lemons, etc.) were present before the viewer rather than in some mythical realm.

The range of styles in fine art is immensely greater than what appears in natural photographs and videos that command the attention of traditional computer vision and artificial intelligence research. Furthermore style affects a wide range of sub-problems in computational art analysis, including semantic segmentation, object recognition, image summarization, semantic analysis, and question answering. Artificial intelligence research has yet to adequately address the deep problems of characterizing styles in art and computing how style influences the meaning of an artwork [40].

Small Data Sets The ascendancy of deep-neural-network approaches to problems in image analysis relies upon training such networks with enormous data sets, for instance of images and associated text [15]. It is not uncommon to train networks with hundreds of millions of images, or even more than a billion images, generally scraped from the internet [31]. By contrast, the number of art images—say in the Western canon—is far smaller, perhaps a few million works, depending upon one’s definition of “canon.” Even some of the most prolific artists left us relatively few works. For instance perhaps the most-prolific major artist, Pablo Picasso, left us roughly 13,500 paintings and Rembrandt roughly 600. At the other extreme are important artists such as Johannes Vermeer, who left us only about 34 paintings, and Leonardo just 19.

Given the great variety of styles and subjects and deeper semantic analysis challenges in the Western canon, we might expect that we will need datasets proportionally *larger* than are typically used in research on photographs. In short, the relatively small size of art databases is a central limitation in many computational approaches to art analysis, including the areas listed here.

Development of an Artwork Revealed by Multiple-Mode Imaging X-radiography and infrared reflectography reveal underdrawings, pentimenti, and

ghost paintings in many artworks. Such imaging provides vital information about how an artist developed an artwork and thus how the work should be interpreted. More exotic imaging, such as Raman Spectroscopic Imaging, reveals atomic elements throughout a painting, including in (hidden) underdrawings that otherwise appear achromatic through x-radiography and infrared reflectography [5]. Such atomic elements often indicate the presence of particular pigments, such as Cadmium red, Lead and Zinc white, Copper blue, and others. While traditional computer vision addresses multispectral imaging (for instance in remote sensing), it rarely if ever analyzes the essential *differences* in images, including overall designs, in the service of interpreting the work as a whole.

Thus there are great opportunities for computer vision and artificial intelligence to use information from a range of imaging modalities to aid in the interpretation of fine art paintings, including issues such as authentication and attribution.

Non-realistic Objects Much artwork in the Western canon is based on artists' imaginations, myths from the Bible or Ancient Greek and Roman literature, and so forth. The resulting images need not be tethered to the constraints and objects that appear in the natural world. Thus dragons, gryphons, angels, chimera, centaurs, fantastical sea serpents, and more appear throughout the canon. Some non-realistic objects may be amalgams of portions of common objects or creatures (as for instance fauns and satyrs in Roman mythology are a blend of a goat and a man), and thus might be rather easily learned by deep neural networks while others, such as Hecatoncheires from Greek mythology, would likely be far more difficult. The import of the above for artificial intelligence approaches to art research is that the required large training sets, almost universally taken from natural photographs, simply will not include depictions of a significant number of such rare or occasionally unique objects which are almost always essential to the interpretation of an artwork in which they appear.

Unphysical Conventions The vast majority of natural photographs obey constraints of the natural world such as geometric perspective, foreshortening, illumination from above, consistent shadows, and so on. Artists are not bound by such constraints, of course, and indeed much of art—notably Surrealist, Expressionist, and mystical religious work—deliberately flaunt such conventions in service of aesthetic, artistic, social, or philosophical ends. With rare exceptions, traditional computer vision and artificial intelligence have not addressed the analysis and interpretation of images that break these conditions with the rare exceptions being image analysis of non-realistic computer graphics worlds [35]. Given the extraordinary wealth of non-realistic and unphysical representations found in art, it appears likely that either vast databases of such art, or improvements in neural architectures, or both, will be needed to computationally analyze non-realistic representational art.

Abstraction An extremely important and diverse genre of Western art is Abstractionism where, by definition, the images do not represent traditional objects. Abstraction spans a range as broad and diverse as the works of Piet Mondrian to

Jackson Pollock to Frank Stella to Wassily Kandinsky to many others. For this reason, the vast majority of natural photographs are of little or no relevance to the analysis of abstract art. Even natural photographs of rather abstract images (mottled reflections from a water surface, closeups of textured surfaces such as leaves or crushed seashells) will differ markedly from the kind of deliberate abstractions that appear in styled artworks executed in paint.

Abstract art nevertheless presents a number of image analysis problems not addressed by mainstream artificial intelligence research, most notably authentication [22]. Computational higher cognitive interpretation of abstract art is a challenge of the highest order, and will likely resist significant successes for a very long time. Regardless, we can be confident that computational analysis of abstract works will profit from large databases of artworks in characteristic abstract styles.

Authentication and Attribution The problem of fake and forged art—works produced or presented so as to deceive the viewer—is profound and extensive; it has great import to both academic scholarship and the commercial art markets [7, 20, 34]. Even the related case of authentication of a work, where no deception is involved, can be extremely complex [44]. With rare exceptions, traditional artificial intelligence has hardly addressed the problem of authorship in its purest form. After all, authorship is rarely of concern in natural photographs. One related task that has received significant research attention is that of detecting of deep fake photographs [2]. However, most of the mainstream approaches to this problem have little or no relevance to fine art authentication studies [13].

Recovery of Lost Art The lost artworks *that we know of* would fill the walls of every one of the world’s public museums [8]. Art is lost for many reasons, including fire, flood, war, theft, iconoclasm, vandalism, and more. Recent approaches to art recovery view the problem as one of finding an estimate of the missing work most consistent with surviving information of various forms, such as copy works, preparatory sketches and cartoons, textual and verbal descriptions, surviving works by the artist (author) in question, and more. Traditional artificial intelligence research has addressed the creation of works *in the style* of a particular author and based on a textual description (such as “A portrait of a young woman on a bicycle in the style of Johannes Vermeer”), but only recently has work sought to incorporate a broader range of textual and image information into the recovery of missing works [12].

Semantics Perhaps the most important class of problems posed by art analysis that is not currently addressed by traditional computer vision and artificial intelligence research centers on *meaning*, or more properly stated, *semantics*. Mainstream image analysis seeks to compute descriptions of images, or to answer questions about the depicted objects and their relations, for instance describing a pose or action [25]. Semantic analysis of fine art is rather different, however, and is a much harder problem. Here the task is to infer a message or meaning crafted by the artist (called the “author” in these contexts); such meanings are far more tenuously linked to the image per se. Thus for instance vanitas still lives of the Dutch Golden Age

depict skulls, rotting flowers and fruit, recently snuffed candles, books, musical instruments, pocket watches, and so on, as each of these objects has a well-understood meaning, all working together to create a “visual sermon.” The primary (religious) message or meaning of such works might be summarized as:

Live a humble, sober, virtuous life in preparation for the eternal life to come.

Here, too, the color palette, style, composition, and formal properties of the work (e.g., it being life-sized) all contribute to the meaning [26].

Again, traditional semantic image analysis might merely describe the depicted objects and their relations yet not consider the central problem of how the objects, the composition, and style work together to convey a meaning or message crafted by the author [40].

Several of the above challenges are interrelated. For example, automatic analysis of non-realistic tableaus is closely related to the problem of small data sets. After all, if there were sufficiently large datasets of artworks bearing non-realistic tableaux presumably strong learning systems such as deep neural networks would be able to learn the regularities in such tableaus. Likewise, computational authentication studies will benefit from large data sets, much as accuracy in pattern classification improves the larger the representative training set [10].

The single problem that relates to the largest set of listed areas is that of small data sets. For that reason, we consider a method for compensating for small data sets associated with art analysis.

20.2.2 Traditional Approaches to Learning with Small Data Sets

A common yet very poorly understood cognitive ability of normal humans is one-shot (or few-shots) learning, where the subject can learn a pattern, category, or visual concept from just one or a few visual presentations of exemplars [27]. In the case of cognition of art, a subject can learn the style of say Jackson Pollock or Piet Mondrian or numerous other artists from just one or a few of their artworks. The neural mechanisms subserving such learning strongly suggest that the relevant processing is at a rather late, cognitive level, in which lower-level features and attributes are integrated to form a coherent novel concept [28, 29].

There are deep statistical laws that place limits on the accuracy of learning systems. In particular, the so-called *bias-variance tradeoff* states that under extremely broad conditions powerful (expressive) learning systems, as are necessary for art analysis, trained with small datasets can at best have low bias or low variance but not both [10, 14, 50]. The only way to get low bias *and* low variance in a highly expressive learning system is to train that system with very large corpora.

There are several approaches to the problem of small data sets in such contexts [21]. One way to address this fundamental tradeoff, and provide an accurate

and reliable learning system, is to restrict the expressiveness of their learning system with prior knowledge of the problem domain. Thus, for example, one can impose symmetries, or a rather general preference toward “simple” models through regularization [17]. One can impose specific prior knowledge of the problem domain as well but this seems rather difficult in most problems in art analysis. After all, how would one impose in a network the prior information that artworks in question depict stories from the Bible?

Another approach is to employ *transfer learning*, where a network is first trained on a separate (but presumably related) task using a large training corpora and then trained with the smaller corpora matching the task at hand. This method can be accurate to the extent that the visual features and statistics relevant to the first task or data set are also relevant to the second task or data set. Thus one might start with a deep network trained on general photographs and then further train it with images of artworks depicting stories from the Bible. It is this approach that we illustrate below.

20.3 The Problem of Small Data Sets

Although learning and optimization has been applied to nearly all tasks in image analysis, we consider for the moment the task of image-based object classification as here the metric of performance (classification accuracy) is particularly clear. In the absence of prior information about the distribution of features in different categories, there is no learning system will, on average, outperform others. Equivalently, under such conditions all learning algorithms perform equally well (or poorly), on average [46, 47].

Given a classifier of a given complexity or expressivity (as expressed, informally, by the number of free parameters or Vapnik-Chervonenkis dimension [1]), the number of training patterns should scale as the number of free parameters in the learning system. An informal guide for early three-layer neural network systems is that the number of training patterns should be roughly ten times the number of weights in the network. Of course this is a heuristic, a mere guideline which derives from the classes of problems addressed rather than some deep theoretical analysis.

20.3.1 One-Shot Learning

In many domains, humans can perform one-shot (or few-shots) learning, that is, learn a category or task with just one or a few exemplars [27]. In our current context, such one-shot learning includes the ability to learn the style (for example color, brush stroke, compositional principles, and so on) of a painter by just one training exemplar. Most observers unfamiliar with fine art can nevertheless learn the distinctive styles of painters such as Mark Rothko, Piet Mondrian, J. M. W. Turner,

Claude Monet, Ellsworth Kelly, Jackson Pollock, and others from such a single presentation. The neural mechanisms subserving such one-shot learning are rather incompletely understood, but it seems that visual novelty detection and associated orienting responses are involved [29].

Two of the leading approaches to one-shot learning in deep neural networks include multiple presentations of the single exemplar, or training with an unusually high learning rate, or storing single “novel” exemplars [4]. There has been preliminary work on computational models of one-shot learning for image segmentation, our example problem, below [37]. These approaches to one- or few-shots learning have yet to perform at the level of accuracy needed for complex art analysis, however.

A common general approach, then, is to incorporate directly into the learning system prior knowledge about the problem domain. However, this approach becomes increasingly difficult for high-level image analysis of art by deep neural networks, such as image summarization, where information is distributed through many millions of connection weights and parameters [15]. Even if the system designer has prior knowledge—that the images in question all include animals, say—it is extremely difficult to incorporate such knowledge into the network architecture directly. Our approach, then, is to incorporate such knowledge *indirectly*, as we shall see.

20.3.2 Small Data Sets in Art Analysis

The bias-variance tradeoff places rigorous statistical limits on the accuracy (or bias) and variability (or variance) of learning systems and almost always the larger the training set of representative exemplars, the more accurate the trained system [10, 18]. As we just saw, there simply are not enough art images, taken alone, for training deep networks for high-level art analysis. We can incorporate prior knowledge *indirectly* by so-called *manufacturing data*, where in our current context we call such data *surrogate artworks* or surrogate art images, as we discuss below.

The precise number of paintings in the Western canon depends upon one’s definition of the canon itself, of course, but we can make a rough estimate of the number of works as follows. An extensive quantitative analysis of art trends in nineteenth-century Western two-dimensional art considered 500,000 artworks [16]. We can informally assume that there were fewer works in the canon in each of the earlier centuries, and likely more works in the twentieth century. A very rough estimate, then, would be for the last ten centuries perhaps $5M$ works.

Thus the number of works in the Western canon is a small fraction of the hundreds of millions or even several billion photographs used to train state-of-the-art deep neural networks. Note, too, that the stylistic variety in paintings is vastly greater than we find in photographs. Thus for “equivalent” performance on tasks such as object recognition, segmentation, image summarization, and so on, we likely

need yet more paintings than photographs. The problem is even more severe when we focus on the work of a *single* artist. Even the most prolific major artist, Pablo Picasso, left us just 13,500 two-dimensional works; again, a small fraction of the number of photographs used for developing accurate deep network systems.

These admittedly rough estimates suffice to show that small data sets are a clear problem in developing and training deep neural networks for high-level art analysis.

20.4 Style Transfer by Deep Neural Networks

Our general approach to increasing the effective number of training samples is to create a large set of *surrogate* art images by computationally transferring the style from a single artwork or a set of artworks to individual natural photographs, as illustrated in Fig. 20.1, where here a modern portrait photograph is the content image, labelled **p**, and a portrait by Paul Cézanne is the style image, labelled **a** [33]. The result of computational style transfer, is at the right and labelled **x**. Style transfer maps the style (statistics of color, low-level form, etc.) to a content image, and appears in some mobile device apps as amusements,

State-of-the-art style transfer is done by deep neural networks. Consider a deep network, such as VGG-19 trained for object recognition, when presented with an image. Generally speaking, the *style*, as represented in the statistics of color and local form (such as brush strokes), is represented in the *early* layers of the network



Fig. 20.1 In basic style transfer, a *content* image, **p** (here the personal portrait photo), and a *style* image (or set of images), here Paul Cézanne's *Portrait of Uncle Dominique in a turban*, oil on canvas (1866), **a**, are presented to a trained deep neural network. The image then produced by the network, **x**, obeys Eq. 20.1—a weighted balance of matching content and style. We call **x** a “surrogate” art image, which can be used to train deep networks for a number of high-level art-analysis tasks, as discussed in Sect. 20.5

while the *content*, as encoded in compressed representations of objects and their relations, is represented in the *later* layers of the network. Here we can consider $S(\mathbf{x})$ a vector representation of an input image \mathbf{x} at early (style) layers, and $C(\mathbf{x})$ the vector representation of that image at late (content) layers. The details of these two representations are described more thoroughly elsewhere [30].

Computational style transfer can be expressed as an optimization procedure seeking an image, represented as a vector of pixels values, \mathbf{x} , such that the content of \mathbf{x} matches that of the content training images as closely as possible, that is $C(\mathbf{x}) \approx C(\mathbf{p})$, and the style of \mathbf{x} matches that of the style of a single or set of style training images, that is, $S(\mathbf{x}) \approx S\mathbf{a}$. We express the cost or merit function $\mathcal{L}(\mathbf{x})$ as:

$$\mathcal{L}(\mathbf{x}) = |C(\mathbf{x}) - C(\mathbf{p})| + k|S(\mathbf{x}) - S(\mathbf{a})|, \quad (20.1)$$

where $|\cdot|$ denotes the L_2 norm in the pixel space of the style and the content sets of layers in the network. Here k is a scalar that adjusts the relative effective contributions of the content and the style image information: When k is large, the style of \mathbf{x} matches closely that of \mathbf{p} and when k is small \mathbf{x} matches closely that of \mathbf{p} , in an L_2 norm. Note that $C(\mathbf{x})$ and $S(\mathbf{x})$ are typically of different dimensionality as the (style) vectors in the early stages of a network involve more neurons than do the (content) vectors in the late, compressed stages of a network [11]. (We note in passing that one can generalize Eq. 20.1 to implement a blending of *several* styles [30].)

20.5 Semantic Segmentation

We shall illustrate the power of surrogate art data in the high-level task of semantic image segmentation. Segmentation, sometimes called *pixel classification*, is the computational task of assigning to each pixel in an image a category label. Recall that a segmented image resembles a jigsaw puzzle, in which each pixel is classified (and often pseudocolored) the same as other pixels that are sufficiently similar, as shown in Fig. 20.2. Similarity is most frequently computed based on color, but it might also be based on local image statistics or related computable functions at each point. *Semantic* segmentation goes one step further and assigns a category label to each of the segmented regions, for instance PERSON, WATER, SKY, ROAD, and so forth.

Deep networks for semantic segmentation resemble those their non-semantic counterparts but are trained with category labels assigned to each output pixel [3]. Standard supervised learning in such networks requires millions of training patterns to achieve an accuracy high enough to be of use in image analysis. It is costly to hand label segmentation regions in every photograph in such a large database, however, so nearly all systems are created through semi-supervised learning. In semi-supervised learning for image segmentation, regions are hand marked and labeled in a small set of images which are then used to train the deep network. The performance of the

Fig. 20.2 A computational segmentation of Henri Matisse's *The dance* based on color. The pseudocolors displayed are unrelated to the colors in the original painting. This is simple segmentation (not semantic segmentation), so the regions are not labeled



network at this stage is better than chance but not sufficiently high for the full image analysis application under consideration.

Next, unlabeled images from a large corpus are presented, one by one, to the network and segmented by the network, thereby leading to a large corpus of segmented (and labeled) output images. Next the segmentation network is trained with this large newly labeled dataset. In this way, the final trained network achieves high accuracy on the task at hand [6].

20.6 The Value of Surrogate Art Images for Training Deep Networks for Art Analysis

As we shall see below, a deep network fully trained on natural photographs by semi-supervised learning does not perform segmentation of fine art images with high accuracy. The statistics of colors and contours of art images, including of portraits, differ significantly from those of natural photographs and this is surely part of the reason for poor segmentation performance.

In order to train a deep network for accurate segmentation of art images we employ *transfer learning* [32]. Recall that in transfer learning, a network fully trained on one task is further trained with additional patterns on a second task. In our case, the first task is segmentation of natural photographs and the second task—our ultimate goal—is segmentation of fine art images.

Figure 20.3 shows representative segmentation results for a portrait by Pierre-Auguste Renoir [19]. Panel (b) shows the semantic segmentation (HUMAN) performed by a deep network trained with only natural photographs in the COCO dataset [31]. This segmentation is rather poor, as quantified by the *IoU* metric, defined below. Much of the girl's hair is misclassified, presumably because its lightness and texture of the brush strokes resembles that of the background.

Panel (c) shows the segmentation when the original deep network is transfer trained with a small number of art images. Note that the segmentation accuracy is

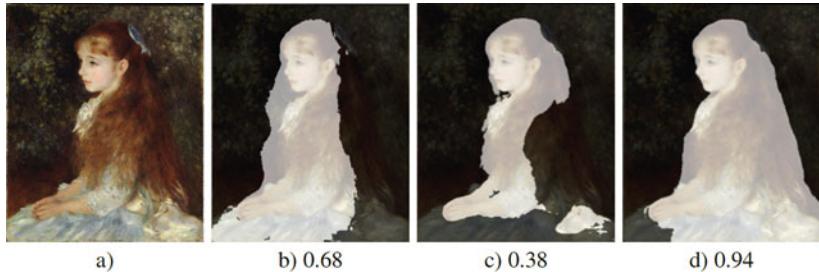
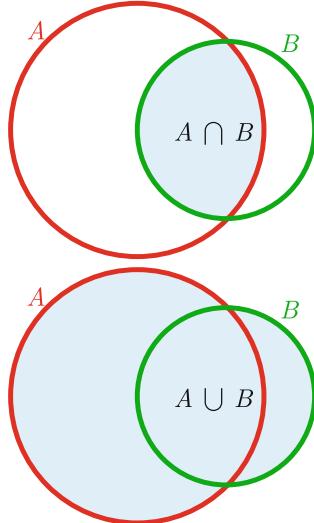


Fig. 20.3 (a) Pierre-Auguste Renoir’s *Portrait of Mademoiselle Iréne Cahen d’Anvers* (65×54 cm), oil on canvas (c. 1880) (Emil Bühl Collection, on long term loan at Kunsthaus Zürich). (b) Segmentation by a network trained on only natural photographs, (c) a network trained with a small number of artworks, and (d) a network trained with a large number of artworks. The segmentation performance, shown by the *IoU* metric reveals that the deep network trained on representative Impressionist paintings performs quite well on this painting. (Figure from [45], which also contains details of the deep network, training set sizes, and so on.)

Fig. 20.4 The Intersection over Union (or IoU) metric, applied to a ground truth region, A , and a segmentation estimate, B , is defined as $\text{IoU} = \frac{|A \cap B|}{|A \cup B|}$, where $|\cdot|$ represents the size or cardinality of a set. In the case of image segmentation, this cardinality would refer to the number of pixels, for instance of a human figure, marked schematically here in blue



degraded, as is evident by the figure itself and the lower *IoU* value. Apparently the statistical relations of color and form learned from the natural photographs learned by the network were disrupted, yet did not match the statistics of art images. Panel d) shows the segmentation when the original deep network is trained with a large number of surrogate art images. Both the image itself and the high *IoU* metric value confirm that the segmentation is highly accurate. These results are consistent with analyses showing that deep networks for pattern recognition prove robust to some, but not all, changes in the distribution of test patterns [48].

Segmentation accuracy is typically quantified by the *Intersection over Union* or *IoU* metric, as illustrated in Fig. 20.4. In this case A represents a ground-truth

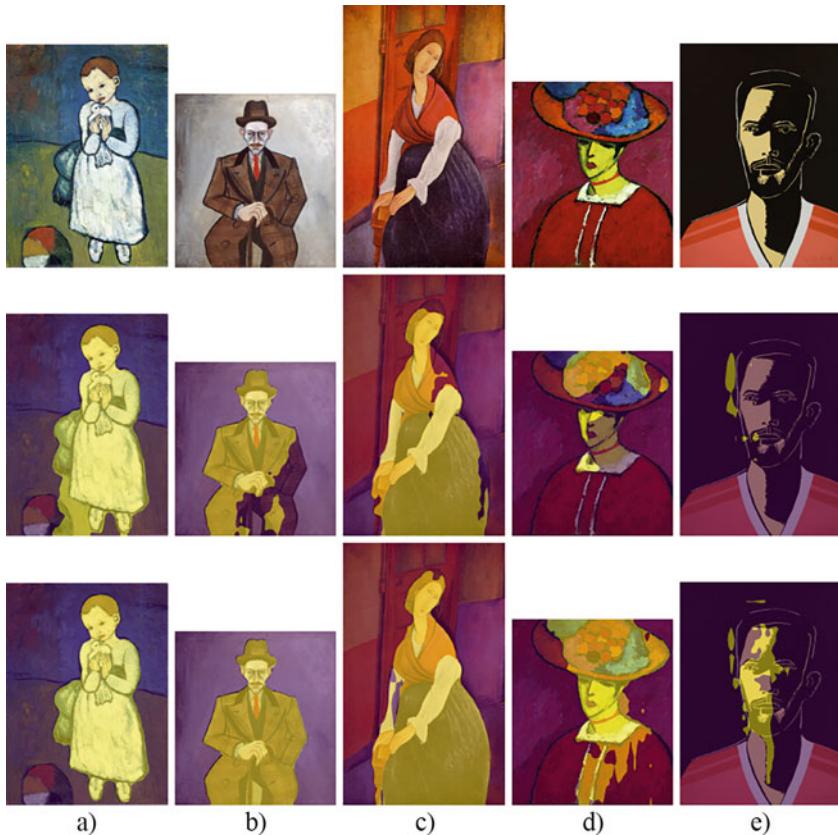


Fig. 20.5 (T) Five portraits, (M) segmentation with a baseline deep network trained on only natural photographs, and (B) segmentation based on a deep network transfer trained with surrogate art images. (a) Pablo Picasso’s *Child with a dove*, (b) Alice Neel’s *Kenneth Dolittle*, (c) Amedeo Modigliani’s *Jeanne Hébuterne in red shawl*, (d) Alexej von Jawlensky’s *Schokko with Wide-Brimmed Hat*, and (e) Alex Katz’s *Daniel*. In all these challenging cases, the network transfer trained with surrogate art images performs segmentation more accurately than the baseline case using a network trained on just natural photographs [19]

segmented region and B the region computed by the segmentation deep network. Of course $0 \leq IoU \leq 1$, where the larger the number the more accurate the segmentation. Note that the only way to have $IoU = 1$ is if $A = B$.

The above analyses and empirical results demonstrate the value of large datasets of art images for high-level computational analysis of fine art images, specifically segmentation. Given, as we have seen, that the number of art images is much smaller than the number of natural photographs used to train deep networks for high-accuracy segmentation of photographs, we asked whether surrogate art images might be used to train such networks. Recall from Sect. 20.4 and Fig. 20.1, above, that a surrogate art image is computed from a natural photograph (the “content”) and

a single or set of art images (the “style”). We showed that a segmentation network trained with such surrogate art images can give high accuracy segmentations—higher than if the network is trained with a large dataset of only natural photographs.

Figure 20.5 shows five portrait paintings and the binary segmentation of PERSON computed by a network trained with just natural photographs (middle row) and by a segmentation network transfer trained by surrogate art images [19]. In these five representative cases, and indeed every example tested, the segmentations from the network trained with surrogate art images was superior to that from the network trained with just art images.

Note that the segmentations of Alex Katz’s *Daniel*, in the right column, were quite poor under both conditions, but that the segmentation derived from surrogate images was nevertheless the superior of the two [19, 45].

20.7 Conclusions

We have seen how the central problem of small data sets in art analysis can be overcome—at least in the limited domain explored—by computational style transfer to form large corpora of *surrogate* art images and then using these surrogate images to train a separate network for the task at hand. This approach led to significant improvement in semantic segmentation of portrait paintings. Presumably this is because the challenges in segmentation are due primarily to the statistics of color and local shape (such as brush strokes) rather than high-level composition and object identities. It is such low-level statistical information that appears in surrogate images, even as the subject matter may not.

There remain numerous high-level tasks in art analysis that seem suitable for the approach described here, including object recognition, image summarization, and possibly authentication and attribution.

This is a set of topics of ongoing research.

Acknowledgments The author would like to thank the Art Libraries and their staffs at Stanford University and the Getty Research Institute, where much of this work was performed.

References

1. Abu-Mostafa, Y.S.: The Vapnik-Chervonenkis dimension: Information versus complexity in learning. *Neural Comput.* **1**(3), 312–317 (1989)
2. Albahar, M., Almalki, J.: DeepFakes: Threats and countermeasures systematic review. *J. Theoret. Appl. Inf. Technol.* **97**(22), 3242–3250 (2019)
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Appl.* **39**(12), 2481–2495 (2017)

4. Bertinetto, L., Henriques, J.F., Valmadre, J., Torr, P., Vedaldi, A.: Learning feed-forward one-shot learners. In: Burges, C.J., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Neural Information Processing Systems, vol. 29, (2016)
5. Bourached, A., Cann, G.H., Griffiths, R., Stork, D.G.: Recovery of underdrawings and ghost-paintings via style transfer by deep convolutional neural networks: A digital tool for art scholars. In: Stork, D.G., Heumiller, K. (eds.) Computer Vision and Analysis of Art. IS&T (2021)
6. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-supervised Learning. MIT Press, Cambridge, MA (2006)
7. Charney, N.: The Art of Forgery: The Minds, Motives and Methods of Master Forgers. Phaidon Press, New York, NY (2015)
8. Charney, N.: The Museum of Lost Art. Phaidon Press, New York, NY (2018)
9. Druzhkov, P.N., Kustikova, V.D.: A survey of deep learning methods and software tools for image classification and object detection. Pattern Recogn. Image Anal. **26**, 9–15 (2016)
10. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. John Wiley and Sons, New York, NY (2001)
11. Dumoulin, V., Shlens, J.S., Kudlur, M.: A learned representation for artistic style (2017). ArXiv.1610.07629
12. Eriksson, J., Bourached, A., Carr, G., Stork, D.G.: Recovering lost artworks by deep neural networks: Motivations, methodology, and proof-of-concept simulations. In: Stork, D.G., Heumiller, K. (eds.) Computer Vision and Analysis of Art. IS&T (2023)
13. Farid, H., Bravo, M.J.: Image forensic analyses that elude the human visual system. In: Memon, N.D., Dittmann, J., Alattar, A.M., Delp, E.J. (eds.) Electronic Imaging: Media Forensics and Security II, vol. 7541, p. 754106. SPIE (2010)
14. Friedman, J.H.: On bias, variance, 0/1-loss, and the curse-of-dimensionality. Data Mining Knowl. Discov. **1**(1), 55–77 (1997)
15. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. In: Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA (2016)
16. Greenwald, D.S.: Paintings by the Numbers: Data-driven Histories of Nineteenth-Century Art. Princeton University Press, Princeton, NJ (2021)
17. Hassibi, B., Stork, D.G.: Second order derivatives for network pruning: Optimal Brain Surgeon. In: Stephen, J.D.C., Hanson, J., Giles, C.L. (eds.) Proceedings of Neural Information Processing Systems, pp. 164–171 (1993)
18. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York, NY (2016)
19. Heitzinger, T., Stork, D.G.: Improving semantic segmentation of fine art images using photographs rendered in a style learned from artworks. In: Stork, D.G., Heumiller, K. (eds.) Computer Vision and Analysis of Art. IS&T (2022)
20. Hoving, T.: False Impressions: The Hunt for Big-time Art Fakes. Simon and Schuster, New York, NY (1996)
21. Iqbal, I., Odesanmi, G.A., Wang, J., Liu, L.: Comparative investigation of learning algorithms for image classification with small dataset. Appl. Artif. Intell. **35**(10), 697–716 (2021)
22. Irfan, M., Stork, D.G.: Multiple visual features for the computer authentication of Jackson Pollock's drip paintings: Beyond box-counting and fractals. In: Niel, K.S., Fofi, D. (eds.) Electronic Imaging: Image processing: Machine Vision Applications II, vol. 7251, pp. 7251Q1–11. SPIE/IS&T, Bellingham, WA (2009)
23. Johnson, Jr., C.R.: Exploiting weave maps. In: Johnson, Jr., C.R., Sethares, W.A. (eds.) Counting Vermeer: Using Weave Maps to Study Vermeer's Canvases, chap. 6. RKD Studies, countingvermeer.rkdmonographs.nl/ (2017)
24. Johnson, Jr., C.R., Hendriks, E., Berezhnoy, I.E., Brevdo, E., Hughes, S.M., Daubechies, I., Li, J., Postma, E., Wang, J.Z.: Image processing for artist identification. IEEE Signal Process. Mag. **25**(4), 37–48 (2008)
25. Khawaja, S.A., Lee, S.L.: Semantic image networks for human action recognition. Int. J. Comput. Vis. **128**(2), 393–419 (2020)

26. Kim, D., Liu, B., Elgammal, A., Mazzone, M.: Finding principal semantics of style in art. In: IEEE International Conference on Semantic Computing (ICSC), vol. 1, pp. 156–163 (2018)
27. Lake, B.M., Salakhutdinov, S., Gross, J., Tenebaum, J.: One-shot learning of simple visual concepts. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 33, no. 33, pp. 2568–2573 (2011)
28. Lake, B.M., Salakhutdinov, R., Tennenbaum, J.B.: One-shot learning by inverting a compositional causal process. In: Burges, C.J., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Neural Information Processing Systems, vol. 26 (2013)
29. Lee, S.W., Doherty, J.P.O., Shimojo, S.: Neural computations mediating one-shot learning in the human brain. *PLoS Biol.* **13**(4), e1002137 (2015)
30. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 386–396. Curran Associates, Red Hook, NY (2017)
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in COntext. In: European Conference on Computer Vision, pp. 740–755 (2014)
32. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: International Conference on Machine Learning, pp. 2208–2217 (2017)
33. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4990–4998 (2017)
34. Moses, N. (ed.): Fakes, Forgeries, and Frauds. Rowman & Littlefield, Lanham, MD (2020)
35. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) European Conference on Computer Vision, pp. 102–118 (2016)
36. Russell, S., Norvig, P. (eds.): Artificial Intelligence: A Modern Approach, 4th edn. Pearson, Hoboken, NJ (2021)
37. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. *CoRR* abs/1709.03410 (2017)
38. Sizyakin, R., Cornelis, B., Meeus, L., Dubois, H., Martens, M., Voronin, V., Pižurica, A.: Crack detection in paintings using convolutional neural networks. *IEEE Access* **8**, 74535–74552 (2020)
39. Stork, D.G.: Optics and realism in Renaissance art. *Sci. Am.* **291**(6), 76–84 (2004)
40. Stork, D.G.: Automatic extraction of meaning in authored images such as artworks: A grand challenge for AI. *ACM Trans. Cult. History Comput.* **15**(4), 1–11 (2022)
41. Stork, D.G.: Pixels & Paintings: Foundations of Computer-assisted Connoisseurship. Wiley, Hoboken, NJ (2023)
42. Stork, D.G., Collins, J., Duarte, M., Furuichi, Y., Kale, D., Kulkarni, A., Robinson, M.D., Tyler, C.W., Schechner, S., Williams, N.: Did early Renaissance painters trace optically projected images? The conclusion of independent scientists, art historians and artists. In: Stanco, F., Battiatto, S., Gallo, G. (eds.) Digital Imaging for Cultural Heritage Preservation, chap. 8, pp. 379–407. CRC Press, Boca Raton, FL (2011)
43. van der Maaten, L., Erdmann, R.G.: Automatic thread-level canvas analysis: A machine-learning approach to analyzing the canvas of paintings. *IEEE Signal Process. Mag.* **32**(4), 38–45 (2015)
44. von Sonnenburg, H.: Rembrandt/Not Rembrandt in the Metropolitan Museum of Art: Aspects of Connoisseurship. Metropolitan Museum of Art, New York, NY (1995)
45. Wödlinger, M., Heitzinger, T., Stork, D.G.: Artist-specific style transfer for deep net semantic segmentation of paintings: The value of large corpora of surrogate artworks. In: Stork, D.G., Heumiller, K. (eds.) Computer Vision and Analysis of Art. IS&T (2022)
46. Wolpert, D.H.: The lack of a priori distinctions between learning algorithms. *Neural Comput.* **8**(7), 1341–1390 (1996)
47. Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**(1), 67–82 (1997)

48. Yamada, Y., Otani, M.: Does robustness on ImageNet transfer to downstream tasks? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9215–9224 (2022)
49. Yuille, A.L., Liu, C.: Deep nets: What have they ever done for vision? *Int. J. Comput. Vis.* **129**, 781–802 (2020)
50. Zhang, Z., Li, J., Stork, D.G., Mansfield, E., Russell, J., Adams, C., Wang, J.Z.: Reducing bias in AI-based analysis of artworks: Overview and tutorial. *IEEE BITS Inf. Theory Mag.* **2**(1), 36–48 (2022)

Chapter 21

Balance of Unity and Variety in Fine Art Paintings: A Computational Study



Jia Li

Abstract A balanced composition is essential for achieving high aesthetics in visual artworks. Seemingly endowed with a natural sense of balance, viewers can often instantly feel the lack of balance in a composition. Yet balance is a rich and abstract notion hard to articulate in quantitative terms. Existing studies of balance in pictures have focused on the spatial arrangement of visual elements, e.g., color and texture. Inspired by the early twentieth century painter Harold Speed's seminal book “The Practice and Science of Drawing”, I investigate in this chapter the balance between two opposing qualities of a picture: unity and variety. An overarching principle implicitly assumed by Speed is that a high level of variety in one visual aspect, e.g., tone, should be compensated by a low level in other visual aspects, e.g., shape, to maintain a proper extent of unity. This work aims at developing computational methods to facilitate the application of Speed's principle in automatic systems. In particular, I have studied the interplay between the variety levels in shape, tone, and color in a painting. First, features to measure variety in shape, tone, and color, were derived using machine learning methods. Then the relationships between these features were examined by linear regression and quantile regression. Because the unity level of a good design varies in an acceptable range, only when the unity level risks being too low do we start to observe a clear tradeoff between the variety levels in different visual aspects. I thus employed quantile regression to model the extreme cases, for instance, paintings with the top variety levels in shape at any given level of variety in tone or color. Experiments on more than 4000 fine art paintings of 34 artists yielded findings that support the design principle of maintaining enough unity by limiting the overall amount of variety in different visual aspects. The techniques developed here enable us to transform the notion of balance advocated by Speed into well-defined quantities. They have potential applications in multimedia information systems, such as providing automatic critiques on the composition of photos or making suggestions for altering an artwork.

J. Li (✉)

Department of Statistics, The Pennsylvania State University, University Park, PA, USA
e-mail: jiali@psu.edu

21.1 Introduction

A pleasant composition of a picture usually appears balanced from various perspectives. Existing studies on the effect of balance on the aesthetic appeal of a painting or photo focus on the spatial arrangement of visual elements in a picture plane [1, 2, 4, 8, 11]. A visual element is often defined as a mass or shape in the picture, which is given a “weight” determined by factors such as size, location, and color. How the weights distribute over the picture plane decides whether the arrangement is balanced [4]. Harold Speed [10] introduced the notion of rhythm to mean the abstract visual power of picture elements, *e.g.*, lines, tones, and colors, separated from the imitation of natural phenomena. According to Speed, an appealing quality in rhythm must be achieved by balancing unity and variety. What is unique about Speed’s discussion is that he views balance in terms of the harmonic coordination of variations in different visual aspects, a standpoint profoundly different from spatial arrangements. I hereby quote Speed for his definitions of unity and variety [10]: *“Unity is concerned with the relationship of all the parts to that oneness of conception that should control every detail of a work of art. All the more profound qualities, the deeper emotional notes, are on this side of the subject. On the other hand, variety holds the secrets of charm, vitality, and the picturesque, it is the ‘dither,’ the play between the larger parts, that makes for life and character.”*

Speed discusses how the senses of unity and variety are conveyed by two types of visual elements: lines and masses. By masses, he means shapes, for instance, regions that are relatively homogeneous in color. In my study, I focus on unity and variety conveyed by the masses. Specifically, regions (aka, segments) in an image generated by a segmentation algorithm are taken as masses. How to define unity or variety in computational terms is a perplexing problem. It is even difficult to define those terms qualitatively. Speed appeals to examples of tonal arrangements in the work of old masters, such as a low-toned sky, to explain how to achieve unity through specific schemes. However, he only identifies a handful of composition schemes that achieve unity but does not describe generally applicable visual characteristics that ensure unity. With deep admiration for Speed’s work, I believe that the way he addresses unity reflects the fact that the sense of unity is deeply rooted in psychology—we can consistently feel unity but have great trouble defining it. As for variety, Speed seems to suggest that higher contrast in tone or color creates the impression of more variety, which also aligns with our heuristics. Again, I quote him: *“In speaking of variety in mass we saw how the nearer these tones are in the scale of values, the more reversed and quiet the impression created, and the further apart or greater the contrast, the more dramatic and intense the effect..... Generally speaking more variety of tone and shape in the masses of your composition is permissible when a smaller range of values is used than when your subject demands strong contrasts..... This principle applies also in the matter of colour. Greater contrasts and variety of colour may be indulged in where the middle range only of tones is used, and where there is little tonal contrast, than where there is great contrast.”*

Inspired by Speed, I constructed a computational framework to study unity and variety as follows. Variety in tone or color is quantified by contrast in either aspect, and variety in shape is quantified by how complex a region appears. Speed has noted that the total amount of variety across the three visual aspects—shape, tone, and color—cannot be too high, or else unity is lost. These three visual aspects are referred to as channels. While acknowledging that my approach to quantifying unity is limited, I restricted the study of unity to a quantity that opposes overall variety across the three channels. According to Speed, unity is also established in ways other than shape complexity, tonal contrast, or color contrast, but it is beyond the scope of this chapter to consider a richer set of visual characteristics.

Equipped with the mathematical definitions of variety in multiple visual channels, I conducted an analysis on over 4000 fine art paintings to test whether the balance of variety exists among shape, tone, and color. This investigation to computationally validate Speed’s principle of unity and variety is motivated by the possibility of mathematically characterizing an abstract design principle and subsequently facilitating the usage of this principle in automatic computer-based systems. Had the analysis not supported the principle, I would have suspected that the failure to capture the notions of unity and variety was the cause rather than questioning the validity of the design principle itself. In addition to being of interest in its own right, the ability to express a design principle in computational terms can potentially lead to new applications. For example, it may become possible to build a computer system that could alert artists if the design principle has been violated and suggest visual qualities that could be improved.

The rest of this chapter is organized as follows. In Sect. 21.2, features are developed to measure the complexity of shape, tone, and color for individual segments within an image. In Sect. 21.3, I describe how to compute a visual significance weight for each segment in an image. These weights are used to compute the overall complexity features for the entire image. Finally, the results and findings of the experiments are presented in Sect. 21.4.

21.2 Variety in Shape, Tone, and Color

I first applied the MS-A3C segmentation algorithm of [6] to each painting. In Fig. 21.6, example images are shown with their segmentation results, in which every segmented region is marked by a distinct color. For every segment in an image, I computed three features called *complexity metrics* to quantify variety in shape, tone, and color respectively and denote them by C_S , C_T , and C_C . Briefly speaking, C_S aims at capturing how complex a shape appears, while C_T and C_C are the levels of contrast with the surrounding area in terms of tone and color respectively. A higher value of tonal (or color) contrast corresponds to higher complexity in the sense of tone (or color). In the meanwhile, a significance weight is computed for each segment by the method of [7]. Suppose there are a total of k segments in an

image with complexity $C_{S,i}$, $C_{T,i}$, and $C_{C,i}$ and significance weight w_i , $i = 1, \dots, k$. The overall shape, tonal, and color complexity metrics are:

$$C_{S,img} = \sum_{i=1}^k w_i C_{S,i}, \quad C_{T,img} = \sum_{i=1}^k w_i C_{T,i}, \quad C_{C,img} = \sum_{i=1}^k w_i C_{C,i}. \quad (21.1)$$

Details about the complexity metrics and the significance weights are presented in Sects. 21.2.1–21.3.

21.2.1 Complexity of Shapes

To define an effective complexity measure for a segmented shape, I took a data-driven machine learning approach. For each segment, three raw features are computed: *percentage of border pixels (PBP)*, *border ratio with respect to a circle (BR)*, *fitness to an ellipse (FE)*. An illustration for these features are provided in Fig. 21.1a. PBP is simply the percentage of pixels in a segment that locate at the border. BR is the ratio between the border length and that of a hypothetical circle

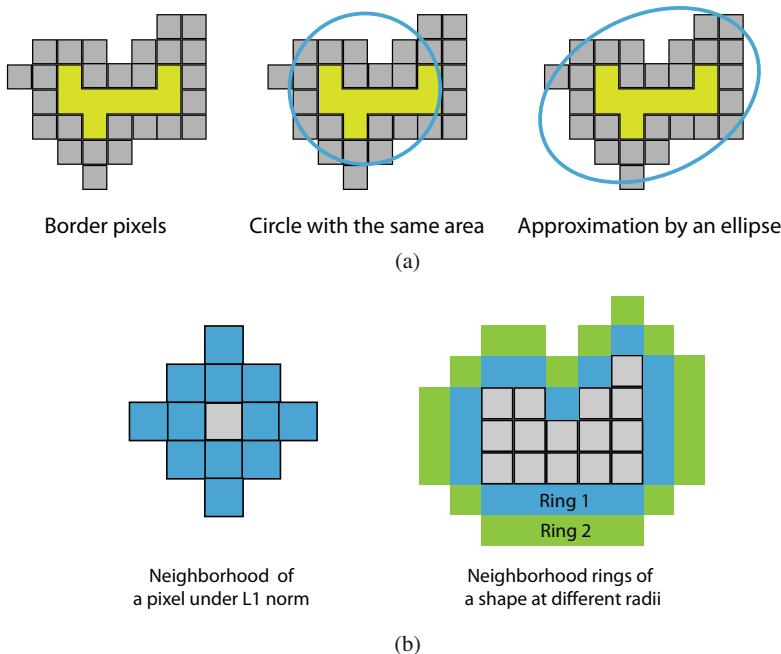


Fig. 21.1 Illustration for features defined. (a) The shape features PBP, BR, FE. (b) How neighborhood rings for a shape are formed

with the same area as the segment. Suppose a segment contains K pixels, among which b pixels are at the border. The area of the segment is defined to be K . A circle with the same area would have radius $\sqrt{K/\pi}$ and perimeter $2\sqrt{\pi K}$. Thus, $BR = b/2\sqrt{\pi K}$. To compute FE, the 2-D horizontal and vertical coordinates of pixels in the segment are treated as a data cloud. Suppose the set of pixels coordinates is $\mathcal{I} = \{(i_1, j_1), \dots, (i_K, j_K)\}$, where (i_k, j_k) contains the horizontal and vertical coordinates of the k th pixel. Let $x_k = (i_k, j_k)^t$. The estimated covariance matrix of the data set $\mathcal{I} = \{x_1, x_2, \dots, x_K\}$ is computed by

$$\widehat{\Sigma} = \frac{\sum_{i=1}^K (x_i - \widehat{\mu})^t (x_i - \widehat{\mu})}{K},$$

where $\widehat{\mu} = \sum_{i=1}^K x_i / K$. Based on the eigenvalue decomposition of $\widehat{\Sigma}$, compute the standard deviation in each of the two principal component directions of \mathcal{I} . Then form an ellipse centered at $\widehat{\mu}$ whose long and short axis are along the first and second principal component directions of $\widehat{\Sigma}$ with respective radii given by twice the standard deviations in the two directions. FE is defined as the percentage of pixels in this ellipse that belong to the segment.

A total of 2984 segments were extracted from 500 paintings. A manual assessment of the complexity of the segmented shapes was conducted. Each segment was shown individually in one image, with the segment highlighted as a flat-colored shape against a flat background. The evaluator assigned each shape with a score of 1, 2, or 3 indicating how complex it appeared. Among the 2984 segments, 1165 were rated 1 (low complexity), 925 were rated 2 (medium complexity), and 894 were rated 3 (high complexity). These scores are considered the values of the response variable, and the three features PBP, BR, and FE as predictor variables. I set the shape type $Y = 0$ (simple) if the score was 1, and $Y = 1$ (complex) if the score was 3. Shapes with scores equal to 2 were not used in training a logistic regression classifier. To summarize, 56.6% of the shapes used in training are of type $Y = 1$, and 43.4% of type 0. Then I estimated a logistic regression classification model based on the data. The posterior probability for being type 1, (highly complex shape) is used as the shape complexity metric. Results show that PBP and FE have little effect on the prediction of Y in the classification model. Hence, in the final model, I only used BR as the predictor variable, which yields an accuracy of 81.54%. Specifically, the shape complexity based on logistic regression is defined as:

$$C_S = \frac{e^{\beta_{S,0} + \beta_{S,1} \cdot BR}}{1 + e^{\beta_{S,0} + \beta_{S,1} \cdot BR}}, \quad (21.2)$$

where the estimated coefficients $\beta_{S,0} = -4.774$, $\beta_{S,1} = 1.763$. The relationship between BR and C_S is shown by the sigmoid function in Fig. 21.2a.

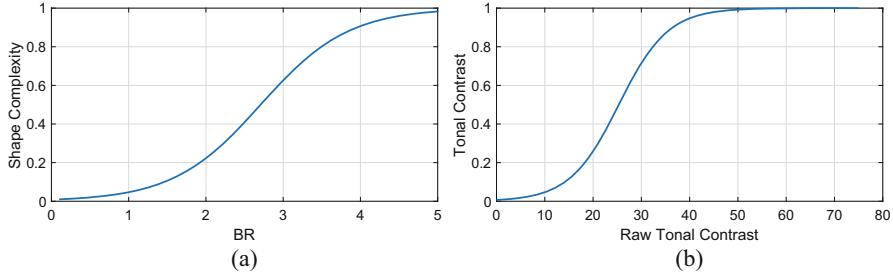


Fig. 21.2 The sigmoid transforms applied to define the shape complexity and tonal contrast of a segment. (a) Shape complexity; (b) Tonal contrast

21.2.2 Tonal Contrast

Based on the terminology of [10], tone refers to the brightness or gray-scale level of a pixel. It is simply computed as the average value of the pixel's Red, Green, and Blue color components in the range [0, 255]. I computed a tonal contrast for each segment with respect to its surrounding area. Suppose the set of K pixel coordinates in a segment is $\mathcal{I} = \{(i_1, j_1), \dots, (i_K, j_K)\}$. Let the tonal value of pixel (i_k, j_k) be I_{i_k, j_k} . As will be discussed shortly, the tonal value is also the intensity component in the HSI color space [3]. The average tonal value of \mathcal{I} is

$$\bar{I} = \sum_{(i,j) \in \mathcal{I}} I_{i,j} / K .$$

Denote the subset of pixels in \mathcal{I} that locate at the border of the segment by $\mathcal{B}(\mathcal{I})$. For a pixel (i, j) , denote its neighborhood under L_1 norm with radius r by $\mathcal{N}((i, j), r)$, which is defined in Eq. (21.3) and illustrated in Fig. 21.1b:

$$\mathcal{N}((i, j), r) = \{(i', j') : |i - i'| + |j - j'| \leq r\} . \quad (21.3)$$

Define the *exclusive neighborhood* of (i, j) with respect to \mathcal{I} as the subset of $\mathcal{N}((i, j), r)$ excluding pixels in \mathcal{I} :

$$\check{\mathcal{N}}((i, j), r) = \mathcal{N}((i, j), r) \sim \mathcal{I} .$$

Then the *neighborhood ring* of segment \mathcal{I} with radius r is defined as:

$$\mathcal{R}(\mathcal{I}, r) = \cup_{(i,j) \in \mathcal{B}(\mathcal{I})} \check{\mathcal{N}}((i, j), r) . \quad (21.4)$$

Neighborhood rings at two different radii for a shape are illustrated in Fig. 21.1b.

The raw tonal contrast based on one neighborhood ring is defined as the average absolute difference between the tonal value of a pixel in the neighborhood ring and the average tonal value of the segment:

$$D_T(\mathcal{I}, r) = \frac{\sum_{(i,j) \in \mathcal{R}(\mathcal{I}, r)} |I_{i,j} - \bar{I}|}{|\mathcal{R}(\mathcal{I}, r)|}$$

To define the raw tonal contrast between a segment and its surrounding area, neighborhood rings at two radii are considered: $\mathcal{R}(\mathcal{I}, r)$ and $\mathcal{R}(\mathcal{I}, 2r)$.

$$\bar{D}_T(\mathcal{I}) = \frac{D_T(\mathcal{I}, r) + D_T(\mathcal{I}, 2r)}{2}.$$

In the experiments, r was set to $\frac{1}{16} \cdot \frac{n_r + n_c}{2}$, where n_r (n_c) is the number of rows (columns) in the image. Next, apply the sigmoid transform to $\bar{D}_T(\mathcal{I})$ to define *tonal contrast*, which varies in the range [0, 1]:

$$C_T = \frac{e^{\beta_{T,0} + \beta_{T,1} \bar{D}_T(\mathcal{I})}}{1 + e^{\beta_{T,0} + \beta_{T,1} \bar{D}_T(\mathcal{I})}}, \quad (21.5)$$

where $\beta_{T,0} = -4.985$ and $\beta_{T,1} = 0.197$. The relationship between \bar{D}_T and C_T is shown by the sigmoid function in Fig. 21.2b.

21.2.3 Color Contrast

To define the color contrast, similarly as with the complexity of shapes, I took a data-driven machine learning approach. The way to define tonal contrast above is not suitable for color because our perception of color depends intricately on multiple dimensions. Besides hue and saturation which are obvious factors affecting color, the tonal value also plays a significant role. For example, in Fig. 21.3a. the four color strips each contain color blocks with the same hue and saturation (computed by the HSI space), while the tone varies from dark to bright. Apparently, our perception of color changes across the strip, suggesting that we cannot contribute color contrast to the chromatic characteristics alone.

I designed an experiment to collect a manual evaluation of color contrast. A computer program randomly generated 2000 pairs of colors. Each pair of colors was used to create two flipped-foreground-background (FFB) images, aptly called a dual pair. Every image contains a flat-colored circle positioned at the center and a background in another color. In the dual image, the colors of the circle and the background are reversed. See Fig. 21.3b for some examples of dual pairs. Two evaluators independently graded the contrast between every pair of colors by viewing the dual images simultaneously. Three possible scores were assigned: 1 for low contrast, 2 for medium, and 3 for high. For each pair of colors, if the average score of the two evaluators is smaller than 2, the type for this pair is defined as $Y = 0$; and if the score is greater than 2, the type $Y = 1$. Among the 2000 color

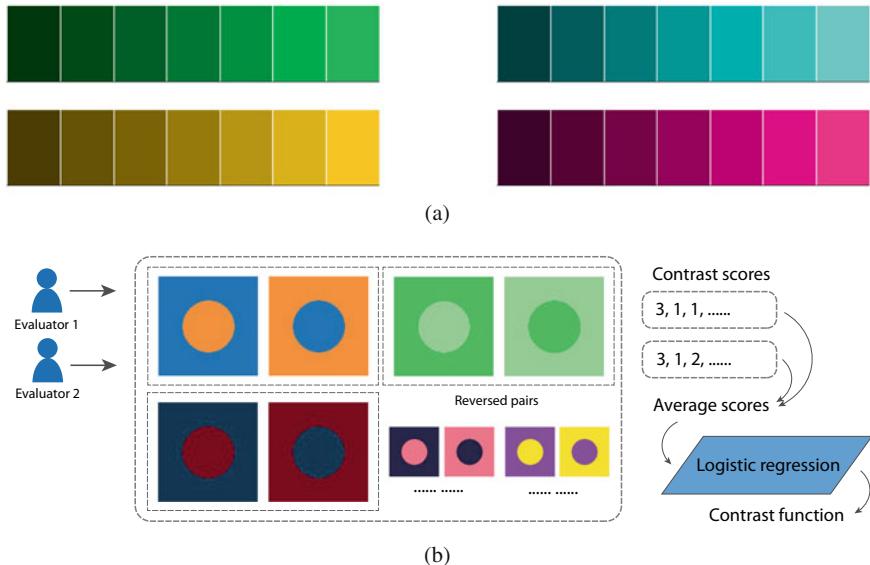


Fig. 21.3 (a) Each strip shows colors with the same chromatic components but varying tonal values. The perception of color is strongly affected by the tonal value in addition to hue and saturation. (b) The process to derive the color contrast function

pairs, 1094 are of type $Y = 0$, 436 are of type $Y = 1$, and the rest 470 were not used in training. The color contrast is defined as the posterior probability for $Y = 1$ based on a trained logistic regression model. Figure 21.3 illustrates the process to derive the color contrast function.

Consider two colors that are originally represented by RGB (Red, Green, Blue) color components. I first converted the RGB color space to the HSI (Hue, Saturation, Intensity) color space. I used a typical HSI color space, with H in the range of [0..360], S in the range of [0, 1], and I in the range of [0, 255]. The formulas for the color space conversion are as follows:

$$\tilde{H} = \cos^{-1} \left[\frac{\frac{1}{2} \cdot ((R - G) + (R - B))}{\sqrt{(R - G)^2 + (R - B)(G - B)}} \right]$$

$$H = \begin{cases} \tilde{H} & B \leq G \\ 360 - \tilde{H} & \text{otherwise} \end{cases}$$

$$S = 1 - \frac{3 \cdot \min(R, G, B)}{R + G + B}$$

$$I = \frac{R + G + B}{3}$$

Table 21.1 Features used for predicting color contrast

$X_1 = \frac{ H_1 - H_2 }{360}$	$X_2 = X_1^2$	$X_3 = S_1 - S_2 $	$X_4 = X_3^2$	$X_5 = \frac{ I_1 - I_2 }{255}$
$X_6 = X_5^2$	$X_7 = \frac{f_1(H_1) + f_1(H_2)}{2}$	$X_8 = \frac{S_1 + S_2}{2}$	$X_9 = \frac{I_1 + I_2}{2 \times 255}$	$X_{10} = \frac{f_2(I_1) + f_2(I_2)}{2}$

Suppose the RGB representations of the two colors are converted to HSI space: $Z_1 = (H_1, S_1, I_1)^t$ and $Z_2 = (H_2, S_2, I_2)^t$. To predict the type of color contrast Y , 10 predictor variables are defined. Denote the augmented predictor vector by $X = (1, X_1, \dots, X_{10})^t$. Definitions for the variables in X are listed in Table 21.1. Denote the mapping from Z_1 and Z_2 to X by $X = g(Z_1, Z_2)$. The transforms $f_1(H) \in [0, 1]$ and $f_2(I) \in [0, 1]$ are defined by

$$f_1(H) = \frac{1}{2} \left[1 + \cos \left(\frac{H}{360} \cdot 2\pi \right) \right], \quad f_2(I) = \frac{1}{2} \left[1 - \cos \left(\frac{I}{255} \cdot 2\pi \right) \right].$$

The logistic regression model fitted has parameters

$$\beta_C^t = (-25.2, 54.56, -57.78, 2.35, -4.25, -3.82, 18.45, 7.68, 5.38, 6.63, 4.18).$$

The posterior probability of $Y = 1$ based on the logistic regression model is

$$P(Y = 1 | X) = \frac{e^{\beta_C^t \cdot X}}{1 + e^{\beta_C^t \cdot X}}.$$

Thus the color contrast between two HSI color vectors Z_1 and Z_2 is defined by

$$\tilde{C}_{HSI}(Z_1, Z_2) \triangleq \frac{e^{\beta_C^t \cdot g(Z_1, Z_2)}}{1 + e^{\beta_C^t \cdot g(Z_1, Z_2)}}.$$

Similarly as with the definition of tonal contrast, to compute the color contrast of a segment and its surrounding area, two neighborhood rings of the segment \mathcal{I} are formed: $\mathcal{R}(\mathcal{I}, r)$, $\mathcal{R}(\mathcal{I}, 2r)$. Suppose the HSI color vector of pixel (i, j) is $Z_{i,j}$, and the HSI of the average color of segment \mathcal{I} is \bar{Z} . The color contrast for segment \mathcal{I} is defined by

$$C_C \triangleq \frac{1}{2} \left[\frac{\sum_{(i,j) \in \mathcal{R}(\mathcal{I}, r)} \tilde{C}_{HSI}(\bar{Z}, Z_{i,j})}{|\mathcal{R}(\mathcal{I}, r)|} + \frac{\sum_{(i,j) \in \mathcal{R}(\mathcal{I}, 2r)} \tilde{C}_{HSI}(\bar{Z}, Z_{i,j})}{|\mathcal{R}(\mathcal{I}, 2r)|} \right] \quad (21.6)$$

21.3 Visual Significance of Segments

A complexity metric for an overall image is defined as the weighted average of the complexity metrics for individual segments. Now I describe how the segment

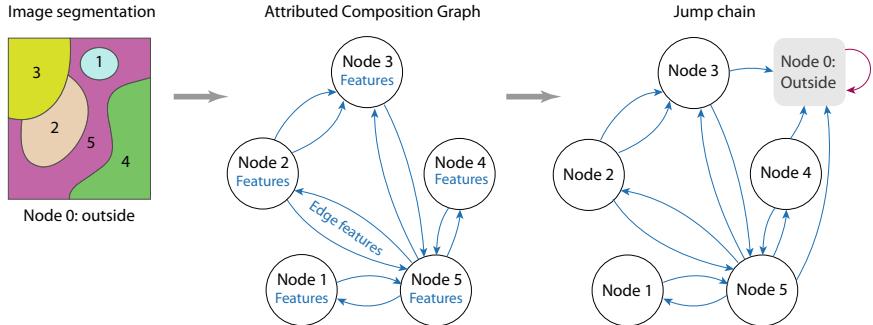


Fig. 21.4 The construction of ACG and the formation of the MC used to model the visual significance of every segment in an image

weights are determined. I first constructed a graph called *attributed composition graph (ACG)* with each node representing one segment. Two nodes are connected if the segments are adjacent in the image plane, that is, one segment is at the boundary of another and vice versa. Three types of attributes are generated for each node:

1. *Visual features for the individual characteristics of a segment*: Some of the features are computed using only information about the pixels in this segment, e.g., average color components in the HSI color space. However, some features depend on the overall composition of the image. For example, the average distance (specifically, the geometric distance in the image plane) of a pixel in a segment to the composition center of the image is included as a feature. To decide the composition center of an image, the center of every segment is a candidate. Among those segment centers, the one that is on average closest to all the other segment centers is chosen.
2. *Visual features capturing the pairwise relationship between the segment and its neighbors*: As such features are associated with two adjacent segments, they are defined on every edge connecting two nodes. I call them *edge features*. Examples include a measure of difference in the orientations of two segments and a measure of the extent that one segment is more visually attractive than the other in terms of color. The former is symmetric between two nodes, while the latter is not.
3. *Visual features characterizing global traits of a segment with respect to the whole image*: Examples include whether the segment plays a special role in the composition of the whole image, e.g., oriented diagonally with respect to the image frame.

To contrast with edge features, I call the first and third types features *node features*. Figure 21.4 illustrates ACG and the jump chain of the continuous Markov chain constructed based on the segmented image.

The main idea of [7] for assessing the visual significance of a segment in an image is to model the time the eye will stay on this segment. The idea was inspired by the opinion of artists that a successful painting should be able to retain a viewer's

gaze inside the image. The composition of a picture is crucial for holding the attention of the viewer. The shift of visual attention is modeled by a continuous time Markov chain (CT-MC), where each segment in the image corresponds to one state of the MC and the area outside the image corresponds to an additional state. A CT-MC is a discrete-time MC (DT-MC) embedded in continuous time; and the waiting time for a state transition to occur follows the exponential distribution with a parameter determined by the current state. I refer to [9] for an introduction to MC. This embedded DT-MC is called the *jump chain*. The spatial arrangement of the segments matters most for setting up the jump chain, while the distribution of the waiting time for a state transition depends mainly on the characteristics of the individual segments. As will be explained shortly, the parameters of the CT-MC are derived from the ACG constructed from the segments in the image.

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where \mathcal{V} is the set of nodes and \mathcal{E} the set of edges. \mathcal{G} is formed by augmenting the graph topology of ACG by an additional node corresponding to “outside image”. The collection of nodes $\mathcal{V} = \{v_0, v_1, \dots, v_k\}$, where v_0 is “outside image” and $v_i, i = 1, \dots, k$ correspond to the k segments of the image. An edge between node i and j exists if segment v_i and v_j are spatially adjacent in the image plane (also called neighbors), as recorded in the ACG. The “outside image” node v_0 is the single absorbing state in the MC with transition probability $P_{0,0} = 1$. Once the MC enters state 0, it will stay in that state forever. Based on the ACG, a weight $e_{i,j}$ is assigned to the edge from i to j . In general, $e_{i,j} \neq e_{j,i}$. The weight $e_{i,j}$ depends mostly on the edge attributes between (i, j) as well as the role of the segments in the overall composition. If no edge exists between i and j , $e_{i,j} = e_{j,i} = 0$. For the jump chain, $P_{i,i} = 0$ for $\forall i \neq 0$. For any segment bordering the outer frame of the image, $P_{i,0} > 0$, otherwise $P_{i,0} = 0$. Between the segments, $P_{i,j} > 0$ if $e_{i,j} > 0$, and $P_{i,j} \propto e_{i,j}$. For each state $i \neq 0$, the time to stay in this state follows the exponential distribution with parameter μ_i , which depends on the node features of state i , e.g., how the contrasts between the segment and its surrounding area and whether the segment appears prominent in the whole image.

The jump chain is not irreducible, and only the “outside image” node v_0 is a recurrent state and all the other states are transient. This means that the MC will eventually enter state v_0 and at that point, the viewing of the image is over—the eye has moved out of the image. Starting from an initial state inside the image, the average amount of visits to each node in the realization of the MC can be computed as follows. Let \mathbf{P}_T be the submatrix of \mathbf{P} that contains the transition probabilities between the transient states. Let $\mathbf{S} = (s_{i,j})$ where $s_{i,j}$ is the expected number of visits which the jump chain spends in state j given that it starts in state i . Let \mathbf{I} be the identity matrix. Then

$$\mathbf{S} = (\mathbf{I} - \mathbf{P}_T)^{-1}.$$

Assume a uniform prior over the transient states for the initial state. Let the expected number of visits the jump chain spends in any state i be s_i and $\mathbf{s} = (s_1, s_2, \dots, s_k)^t$. Denote the vector containing all ones by $\mathbf{1}$. Then $\mathbf{s} = S^t \mathbf{1}/k$.

For a CT-MC, the amount of time the MC spends in a state also depends on the waiting time for a transition to occur at any state. Since the waiting time follows an exponential distribution with rate μ_i , the expected waiting time is $1/\mu_i$. Finally, the amount of time spent in state i is s_i/μ_i , which I define as the *compositional visual significance* (or in short *composition significance*) for state i . A set of normalized significance weights are assigned to each segment i , $i = 1, \dots, k$: $w_i \propto s_i/\mu_i$. Once w_i 's are computed, we can use Eq. (21.1) to compute the complexity metrics in shape, tone, and color for the overall image.

21.4 Experiments

I experimented with 4238 fine art paintings from various schools such as Baroque, Impressionism, and Post-impressionism. Table 21.2 lists the 34 artists and the number of paintings by each artist in the collection. On average, a painting was divided into 10.7 segments, with a standard deviation of 5.3. The histogram of the number of segments per image is shown in Fig. 21.5a. Some example paintings, their segmentation results, and their visual significance maps are provided in Fig. 21.6.

21.4.1 Preliminary Study

As a preliminary study, the correlation coefficients between shape complexity, tonal contrast, and color contrast were computed for the whole collection of paintings and for the paintings of eight individual artists. The results are provided in Table 21.3. Pearson's correlation coefficients between shape complexity and tonal contrast are

Table 21.2 The list of artists and the number of paintings in the dataset by every artist

Artist	Num.	Artist	Num.	Artist	Num.
Milton Avery	20	Giovanni Bellini	31	Rosa Bonheur	53
Jules Breton	77	Caravaggio	74	Paul Cezanne	289
Gustave Courbet	255	Henri-Edmond Cross	20	Paul Gauguin	232
Auguste Herbin	25	Pyotr Konchalovsky	77	Lorenzo Lotto	97
Edouard Manet	106	Ilya Mashkov	52	Henri Matisse	183
Janos Mattis-Teutsch	31	Jean Metzinger	26	Michelangelo	72
Joan Miro	17	Georg Pauli	6	Max Pechstein	27
Pietro Perugino	96	Camille Pissarro	771	Henry Raeburn	58
Rembrandt	324	Guido Reni	123	Luca Signorelli	49
John Singleton	5	George Stubbs	63	Louis Valtat	26
Kees van Dongen	20	Vincent van Gogh	799	Johannes Vermeer	34
Joseph Wright	100				

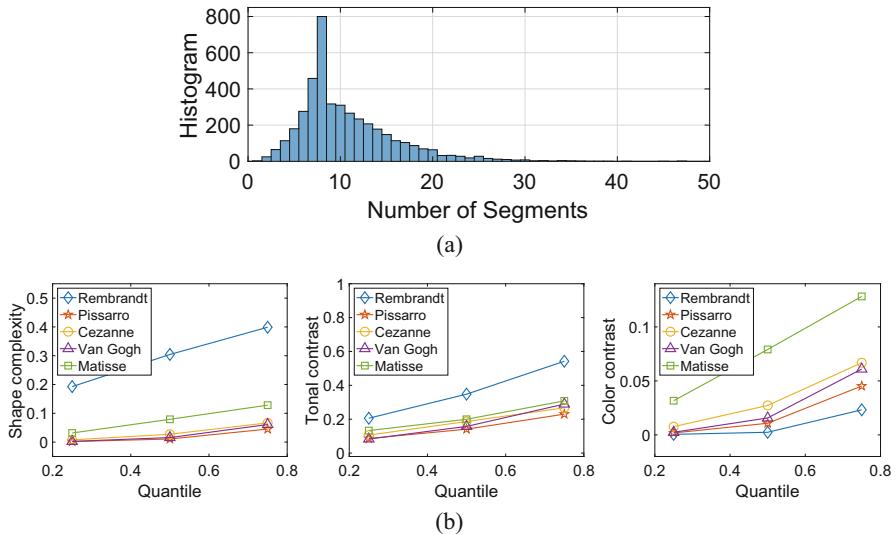


Fig. 21.5 (a) Histogram of the number of segments per image. (b) The 1st, 2nd, and 3rd quartiles of shape complexity, tonal contrast, and color contrast for five artists

always negative, and except for Courbet, the 95% confidence intervals (C.I.) are below zero, indicating that at the significance level 5%, we should reject the null hypothesis that the correlation is zero. Similarly, the correlation between shape complexity and color contrast is negative for all cases. However, for Courbet, Gauguin, and Manet, the 95% C.I.s straddle zero. In contrast, the correlation coefficients between tonal contrast and color contrast are all positive, but five of them are close to zero (below 0.1). As previously explained, it is impossible to separate contrast caused by different pixel values into two independent parts: one resulting from tonal difference and the other from chromatic difference. Our perception of color cannot be fully explained by the chromatic components of HSI, and tonal contrast also contributes to contrast in color. A higher value of tonal contrast tends to be associated with a higher value of color contrast, resulting in a positive correlation between the two types of contrast. For the subsequent analysis, I will therefore only examine the relationship between shape complexity and tonal contrast, color contrast, or both.

Speed's principle of balancing unity and variety emphasizes the importance of maintaining sufficient unity and avoiding too much variety but does not prescribe a fixed level for either. As long as a painting falls within acceptable levels of unity and variety, there is no design requirement for the complexity levels from different perspectives to be related in a specific manner. Thus, I expect that in cases when unity is close to the minimum allowable level or conversely where variety is close to the maximum allowable level, a strong negative correlation between shape complexity and tonal or color contrast will be observed. The linear regression analysis conducted shows the average trend but not the trend existing

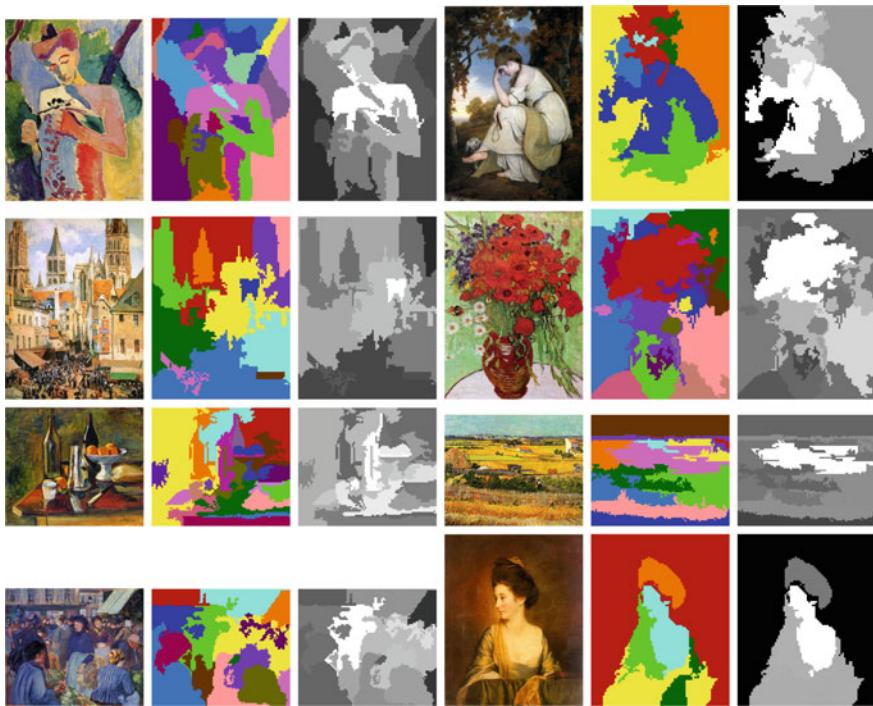


Fig. 21.6 Example paintings, their segmentation results, and visual significance maps. First (or fourth) column: original images. Second (or fifth) column: segmentation results with each segment indicated by a unique color. Third (or sixth) column: visual significance maps. A higher significance value is indicated by a brighter pixel

among extreme cases. The fact that Speed's design principle does not come into play in many artworks explains why the correlation coefficients in Table 21.3 have relatively small absolute values, some of which are not even statistically significant.

21.4.2 Permissible Complexity Based on Quantile Regression

To more accurately understand the relationships between the complexity metrics in the three visual channels, as implied by Speed's principle, I will examine extreme cases using quantile regression and a technique involving moving maximums (to be explained shortly). This analysis will provide insight into patterns present in these cases. A basic assumption here is that when shape complexity, tonal contrast, or color contrast increases, the overall variety of the painting increases, while its unity level drops. Based on Speed's principle, for a painting to be aesthetically pleasing, its unity level must be high enough. Consequently, at any value of tonal contrast or color contrast, the shape complexity should be kept below a certain level, referred to

Table 21.3 Pearson's correlation coefficients between shape complexity, tonal contrast, and color contrast, and their 95% confidence intervals. The coefficients are computed using all the paintings as well as separately using the paintings of individual artists

Artist	Shape/tone	Shape/color	Tone/color
All	-0.30, [-0.33, -0.28]	-0.11, [-0.14, -0.08]	0.05, [0.02, 0.08]
Courbet	-0.05, [-0.17, 0.08]	-0.10, [-0.22, 0.02]	0.09, [-0.03, 0.21]
Cezanne	-0.27, [-0.37, -0.16]	-0.13, [-0.24, -0.01]	0.05, [-0.06, 0.17]
Gauguin	-0.19, [-0.31, -0.07]	-0.11, [-0.23, 0.02]	0.32, [0.20, 0.43]
Manet	-0.33, [-0.49, -0.15]	-0.17, [-0.35, 0.02]	0.42, [0.25, 0.56]
Matisse	-0.15, [-0.29, -0.00]	-0.23, [-0.36, -0.08]	0.04, [-0.11, 0.18]
Pissarro	-0.28, [-0.34, -0.21]	-0.09, [-0.16, -0.02]	0.15, [0.08, 0.22]
Rembrandt	-0.14, [-0.25, -0.03]	-0.11, [-0.21, 0.00]	0.30, [0.20, 0.39]
van Gogh	-0.29, [-0.35, -0.23]	-0.19, [-0.25, -0.12]	0.02, [-0.05, 0.09]

as the *permissible shape complexity*. This value should decrease as tonal contrast or color contrast increases. Unfortunately, we cannot directly observe the permissible shape complexity. Hence I have made two additional assumptions:

1. Except for a small fraction of the paintings, the majority of the paintings created by world-renowned artists adhere to the permissible shape complexity.
2. Because the collection of paintings is large, the permissible shape complexity can be accurately estimated from this collection.

Suppose the percentage of paintings that are believed to exceed the permissible shape complexity is α . I estimated the permissible shape complexity by quantile regression at quantile level $\tau = 1 - \alpha$. How to choose α is inevitably subjective. In Fig. 21.7a, i, based on the whole dataset, the fitted linear quantile regression lines at a series of τ 's are shown, with the shape complexity being the response and the tonal contrast being the predictor variable. For comparison, the linear regression line is also drawn. A similar plot is provided for shape complexity (response) versus color contrast (predictor) in Fig. 21.7b, i. I used the `rq` package in R for quantile regression [5]. As shown by the figures, at any value of τ , the pattern of the quantile regression function is consistent—shape complexity decreasing at the increase of tonal or color contrast. For $\tau = 95\%$ or $\tau = 98\%$, the quantile regression lines differ mainly by a small shift in the intercept. Take $\tau = 98\%$ (i.e., $\alpha = 0.02$), I obtained the following quantile regression functions for shape complexity versus tonal contrast or color contrast:

$$C_s = 0.75 - 0.28C_T , \quad C_s = 0.70 - 0.54C_C .$$

The 98% C.I.s for the two slope coefficients are respectively $[-0.35, -0.22]$ and $[-0.67, -0.40]$.

Quantile regression was also conducted on the paintings of individual artists. Results are shown in Fig. 21.7. These artists were selected because they each have a relatively large number of paintings in the dataset. Also shown by the analysis

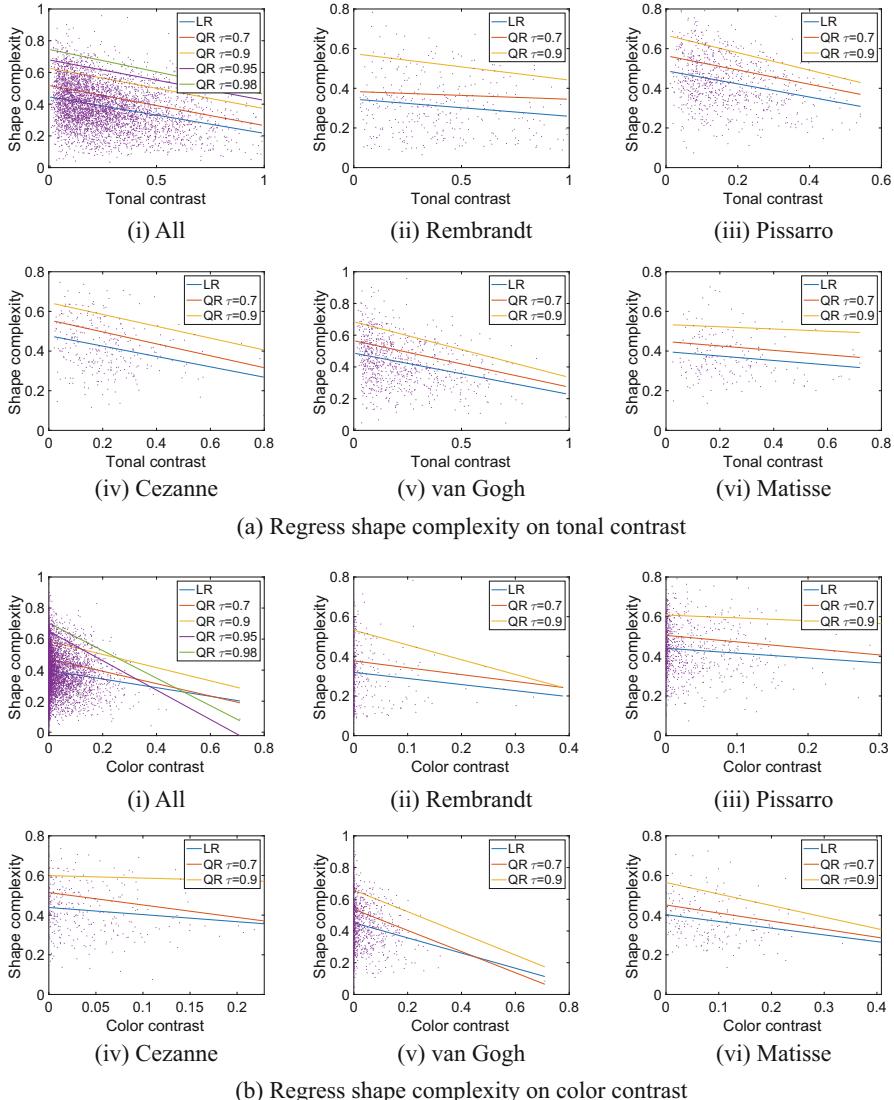


Fig. 21.7 Quantile regression ($\tau = 0.7$ and $\tau = 0.9$) and linear regression lines. **(a)** Shape complexity regressed on tonal contrast. **(b)** Shape complexity regressed on color contrast. In both **(a)** and **(b)**, regression is performed using all the paintings as well as separately using the paintings of five individual artists

below, these artists are interesting to study as some of their paintings possess the maximum shape complexity at a certain level of color or tonal contrast. Since the number of paintings from an individual artist is much smaller than the total number of paintings, it becomes difficult to estimate the quantile regression function when τ

approaches 1. Thus the plots only show results at $\tau = 0.7, 0.9$. In addition, the slope coefficients of the regression lines and their 95% confidence intervals are given in Table 21.4. As shown, all the slope coefficients are negative, consistent with the expectation that the permissible shape complexity decreases when tonal or color contrast increases. For most of the cases, the C.I.s of the slope coefficients are below zero.

21.4.3 Metrics for Variety and Unity

To propose definitions for the variety metric and unity metric of a painting, I allowed 2% of the paintings in the dataset to exceed the permissible level of variety or equivalently to fall below the permissible level of unity. Quantile regression was conducted for shape complexity on tonal contrast and color contrast simultaneously at $\tau = 98\%$. I obtained the regression function

$$C_S = 0.76 - 0.25C_T - 0.58C_C . \quad (21.7)$$

The 98% C.I.s for the slope of C_T and C_C are respectively $[-0.32, -0.17]$ and $[-0.78, -0.38]$. According to the quantile regression function, with probability 98%,

$$C_S \leq 0.76 - 0.25C_T - 0.58C_C .$$

Or equivalently,

$$1.32C_S + 0.33C_T + 0.76C_C \leq 1 . \quad (21.8)$$

I thus define the variety metric V and unity metric U as follows:

$$V \triangleq 1.32C_S + 0.33C_T + 0.76C_C , \quad (21.9)$$

$$U \triangleq \frac{1}{1.32C_S + 0.33C_T + 0.76C_C} . \quad (21.10)$$

By Eq. (21.8), 98% of the paintings satisfy $V \leq 1$ and $U \geq 1$. For both the variety and unity metrics, the permissible level is scaled to 1.

Table 21.4 The slope coefficients of linear regression and quantile regression functions when shape complexity is regressed on tonal contrast or color contrast. The 95% confidence intervals are shown below the coefficients

Artist	Shape/tone				Shape/color			
	LR	QR $\tau = 0.7$	QR $\tau = 0.9$	LR	QR $\tau = 0.7$	QR $\tau = 0.9$	QR $\tau = 0.9$	
All	-0.23 [-0.25, -0.21]	-0.25 [-0.29, -0.21]	-0.26 [-0.30, -0.22]	-0.27 [-0.35, -0.19]	-0.40 [-0.48, -0.32]	-0.42 [-0.48, -0.26]		
Courbet	-0.04 [-0.14, 0.06]	-0.03 [-0.13, 0.12]	-0.04 [-0.20, 0.08]	-0.22 [-0.49, 0.05]	-0.25 [-0.59, 0.08]	-0.41 [-0.65, 0.23]		
Cezanne	-0.26 [-0.37, -0.14]	-0.30 [-0.46, -0.18]	-0.30 [-0.38, -0.17]	-0.36 [-0.68, -0.04]	-0.63 [-0.91, 0.17]	-0.12 [-0.62, 0.31]		
Gauguin	-0.20 [-0.33, -0.07]	-0.25 [-0.37, -0.17]	-0.28 [-0.54, -0.12]	-0.24 [-0.53, 0.05]	-0.22 [-0.71, 0.01]	-0.49 [-0.75, -0.17]		
Manet	-0.28 [-0.43, -0.13]	-0.44 [-0.56, -0.06]	-0.19 [-0.67, -0.10]	-0.35 [-0.74, 0.04]	-0.64 [-0.89, 0.29]	-0.19 [-1.08, 0.39]		
Matisse	-0.11 [-0.22, 0.00]	-0.11 [-0.36, -0.015]	-0.06 [-0.28, 0.04]	-0.34 [-0.56, -0.12]	-0.41 [-0.65, -0.20]	-0.58 [-0.87, 0.23]		
Pissarro	-0.33 [-0.41, -0.25]	-0.36 [-0.50, -0.30]	-0.44 [-0.58, -0.34]	-0.24 [-0.44, -0.04]	-0.33 [-0.58, -0.08]	-0.14 [-0.45, 0.17]		
Rembrandt	-0.09 [-0.15, -0.03]	-0.04 [-0.14, 0.02]	-0.13 [-0.20, -0.02]	-0.31 [-0.62, 0.00]	-0.34 [-0.53, -0.01]	-0.74 [-0.87, 0.32]		
van Gogh	-0.26 [-0.32, -0.20]	-0.29 [-0.37, -0.21]	-0.35 [-0.38, -0.25]	-0.48 [-0.66, -0.30]	-0.67 [-0.84, -0.52]	-0.68 [-0.82, -0.53]		

21.4.4 Examples of High Complexity Based on Three Visual Channels

To study paintings with shape complexity pushed to the extreme, I calculated the *moving maximum (MM)* for shape complexity across different values of tonal contrast or color contrast. Note that a high value of shape complexity refers to considering the three visual channels together. Precisely speaking, these paintings are extreme in terms of their unity level. Take tonal contrast as an example. Windows of width 0.05 are formed over tonal contrast. Start with the window [0, 0.05] and shift it by step size 0.025 each time, creating windows [0, 0.05], [0.025, 0.075], [0.05, 0.1], so on so forth. In each window, identify the maximum shape complexity yielded by paintings whose tonal contrast values are in that window.

In Fig. 21.8a, the moving maximums in the sliding windows are shown by the triangles. The straight line fitted over these moving maximums shows clearly that the moving maximum of shape complexity decreases at the increase of tonal contrast. Similarly, the moving maximums in shape complexity are found across the windows of color contrast. The window size for color contrast is 0.02 with a shifting step size 0.01. The window size for color contrast is set smaller because the spread of color contrast is substantially less than that of tonal contrast. In the shape complexity versus tonal contrast plot, 30 mm paintings are identified; and in the shape complexity versus color contrast plot, 28 mm paintings are identified. These 58 paintings belong to 18 artists. Figure 21.8c, a stacked bar plot, shows the number of MM paintings by each artist according to either tonal contrast or color contrast. It is interesting to see that van Gogh stands out among the artists. According to both tonal and color contrasts, he has the most MM paintings. By color contrast, the MM paintings are mostly from modern-time (the nineteenth century or later) artists. In particular, van Gogh, Pissarro, and Matisse have the highest numbers of paintings among the MM paintings.

Variety and unity are concepts that go beyond any single complexity metric. To illustrate this point, I show in Fig. 21.5b the 1st, 2nd, and 3rd quartiles of shape complexity, tonal contrast, and color contrast for five artists. Neither van Gogh's nor Pissarro's paintings have particularly high values in these statistics for any single complexity metric. Matisse's paintings have relatively high quartile values in color contrast, but not shape complexity or tonal contrast. If we only consider the distribution of a single complexity metric, the paintings of van Gogh and Pissarro (especially van Gogh) do not stand out. However, in terms of variety and unity, which depend on multiple complexity metrics, these two artists have the most examples of paintings that achieve the highest levels of variety. The exceptional quality of their paintings is a result of pushing variety to its limits when considering the complexity metrics of shape, tone, and color together.

As discussed previously, color contrast is intrinsically influenced by tonal contrast. To better understand the relationship between shape complexity and color contrast independently of tone, I experimented with a subset of paintings with tonal contrast controlled in a narrow range. This subset included 424 paintings with tonal

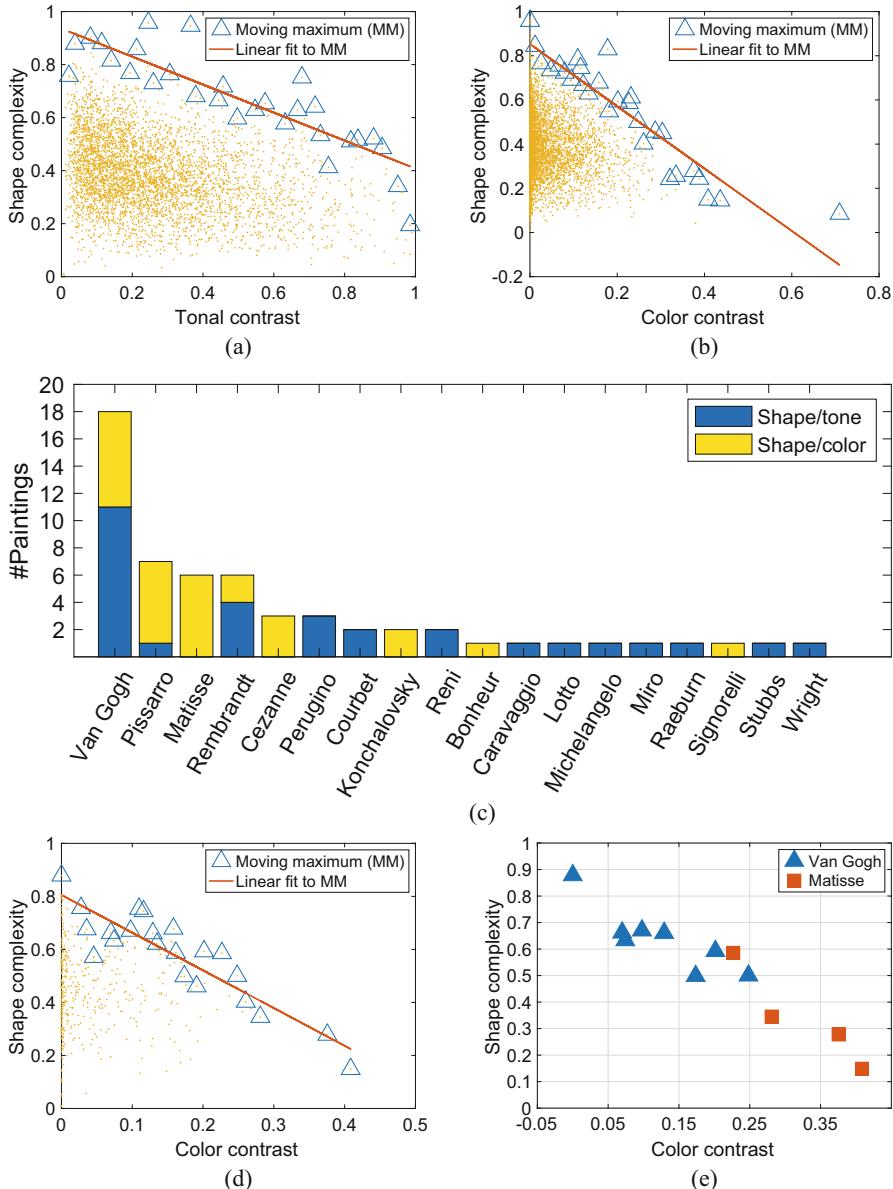


Fig. 21.8 The moving maximums of shape complexity versus (a) tonal contrast and (b) color contrast. (c) The number of paintings by each artist among the paintings that achieve the moving maximums. (d) The moving maximums of shape complexity versus color contrast based on paintings that have the lowest tonal contrast (bottom 10%). (e) van Gogh's and Matisse's paintings among the moving maximum paintings

contrast values in $[0, 0.1]$, which are the lowest 10% in tonal contrast. By using this subset of paintings, I could examine the interplay between complexity metrics of shape and color without the influence of tone. As the tonal contrast is restricted to the lowest end, the color contrast of such a painting comes predominantly from chromatic variation. Similarly, I identified 22 MM paintings, which were by the following artists with the number of paintings given in the parenthesis: van Gogh (8), Matisse (4), Pissarro (3), Cezanne (2), Konchalovsky (2), Mattis-Teutsch (1), Cross (1), Signorelli (1). A detailed plot for van Gogh's and Matisse's paintings is shown in Fig. 21.8e. The paintings of van Gogh distribute in a lower range of color contrast and those of Matisse distribute in a higher range. As shown by the figure, the trend of decreasing shape complexity with increasing color contrast is apparent. Suppose I divide the color contrast into 5 tiers: $[0, 0.05]$, $[0.05, 0.15]$, $[0.15, 0.25]$, $[0.25, 0.35]$, $[0.35, 0.45]$ and the shape complexity into 10 tiers: $[0, 0.1]$, $[0.1, 0.2]$, so on so forth. Figure 21.9 shows the paintings in increasing order of their color contrast values, with the tiers of the color contrast and shape complexity denoted by \mathcal{T}_c and \mathcal{T}_s , respectively, beneath each image. It can be observed that as the color contrast tier increases, the shape contrast tier decreases. The segmentation maps used to calculate shape complexity, shown in Fig. 21.10, reveal that paintings with many details tend to have more intertwined segments, resulting in higher shape complexity. It is interesting to note that in the first four paintings by van Gogh, he primarily used analogous colors with rich, interwoven brushstrokes, resulting in low color contrast but high shape complexity. In contrast, Matisse used complementary colors that form large flat regions, leading to high color contrast and low shape complexity.

21.5 Conclusions

In this chapter, I present methods for defining complexity metrics based on shape, tone, and color using a combination of image processing, computer vision, and machine learning techniques. These metrics have paved a path to computationally validate the composition principle of Speed: unity and variety are balanced by considering the levels of variety expressed in multiple visual channels. I have proposed how to quantify the variety and unity levels of a painting based on quantile regression results. The study includes over 4000 paintings by 34 artists. Among the findings, it is interesting that when examined from a single aspect, such as shape, van Gogh's paintings do not tend towards high complexity compared with the work of other artists. However, if all three visual channels are considered together, many of van Gogh's paintings exhibit extreme levels of complexity while still maintaining unity. Specifically, van Gogh pushed the shape complexity to the highest level when featuring low tonal contrast and wide variations in color contrast.



Fig. 21.9 Paintings that achieve moving maximum. Images are arranged from left to right and top to bottom. The first 8 paintings are by van Gogh, and the last 4 by Matisse. Under each painting, the two values \mathcal{T}_c and \mathcal{T}_s are the tier of color contrast and the tier of shape complexity

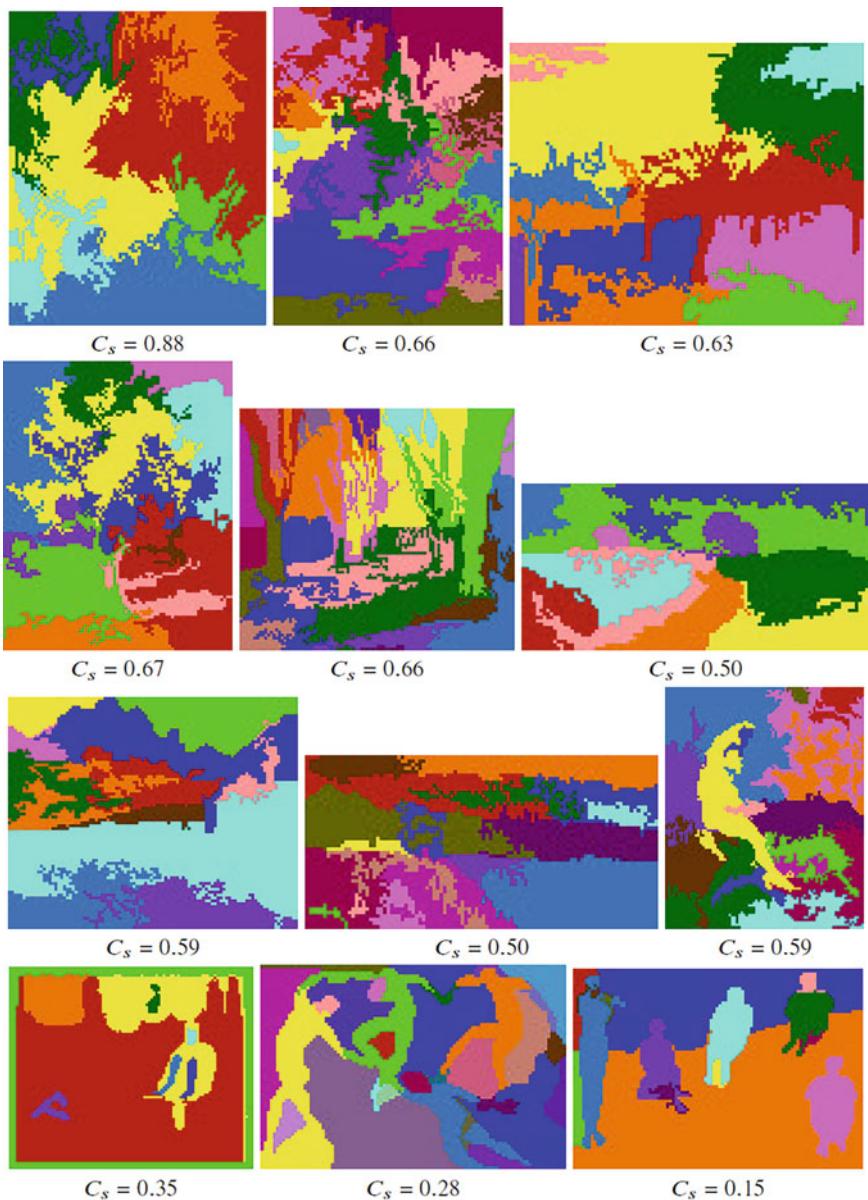


Fig. 21.10 Segmentation results of paintings in Fig. 21.9 and their values of shape complexity

Acknowledgments The author would like to thank Zhuomin Zhang for collecting the painting images from the internet and some of the manual scores for color contrast and for segment shape complexity. The author is also grateful for the reviewers' comments and suggestions.

References

1. Abeln, J., Fresz, L., Amirshahi, S.A., McManus, I.C., Koch, M., Kreysa, H., Redies, C.: Preference for well-balanced saliency in details cropped from photographs. *Front. Hum. Neurosci.* **9**, 704 (2016)
2. Amirshahi, S.A., Hayn-Leichsenring, G.U., Denzler, J., Redies, C.: Evaluating the rule of thirds in photographs and paintings. *Art Percept.* **2**(1–2), 163–182 (2014)
3. Chapanis, A.: Color names for color space. *Am. Sci.* **53**(3), 327–346 (1965)
4. Hübner, R., Fillinger, M.G.: Perceptual balance, stability, and aesthetic appreciation: their relations depend on the picture type. *i-Perception* **10**(3), 2041669519856040 (2019)
5. Koenker, R.: Quantile Regression, vol. 38. Cambridge University Press, Cambridge (2005)
6. Li, J.: Agglomerative connectivity constrained clustering for image segmentation. *Stat. Anal. Data Min. ASA Data Sci. J.* **4**(1), 84–99 (2011)
7. Li, J., Wang, J.Z.: Systems and methods for detection of significant and attractive components in digital images (2019). US Patent 10,186,040
8. Liu, L., Chen, R., Wolf, L., Cohen-Or, D.: Optimizing photo composition. *Comput. Graph. Forum* **29**(2), 469–478 (2010)
9. Ross, S.M.: Introduction to Probability Models. Academic Press, Cambridge (2014)
10. Speed, H.: The Practice and Science of Drawing. Seeley, Service, London (1913)
11. Yao, L., Suryanarayanan, P., Qiao, M., Wang, J.Z., Li, J.: Oscar: on-site composition and aesthetics feedback through exemplars for photographers. *Int. J. Comput. Vis.* **96**(3), 353–383 (2012)

Index

A

- Abstraction
 - in art, 354
- Abstract level, 182
- acknowledgements, xi
- Action Units, 106, 112, 130, 155
 - recognition, 107, 117, 130
- Active Appearance Model (AAM), 130
- Active Shape Model (ASM), 130
- Activism, 95, 97–100
- Aesthetics image datasets, 233
- Affective computing, 128, 332
- Age, 45, 48–50, 52, 53, 147, 153, 335
- Alpers, Svetlana, 220
- Amazon Mechanical Turk (AMT), 333
- Amygdala, 77, 82–85
- Anger, 46, 47, 49–53, 315, 316, 320, 321, 324–327
- Appearance, 151, 152, 154, 156
- Art analysis, 352
- Artificial intelligence, 351
 - in art, 352
 - artificial emotional intelligence, 15
- Artwork
 - development, 353
- Attitude, 5
- Attractiveness, 62, 64–67
- Attributed composition graph, 378
- Attribution, 219, 355
- Audiovisual fusion, 170
- Authentication
 - art, 355

B

- Balance, 198, 369, 381
- Baxandall, Michael, 221
- Bias, 29, 91–95, 97–100, 149, 152, 157
- Bias-variance decomposition, 28
- Bias-variance tradeoff, 356, 358
- Biological motion, 61
- Bodily Expressed Emotion Understanding (BEEU), 165
- Bodily expression, 50, 53, 63, 72, 314
- Body Language Dataset (BoLD), 171
- Body motion, 59–65, 67
- Brain, 46, 72, 73, 78, 79, 84, 85
- Breath, 203–210, 212, 215, 216

C

- Capacity, 27
- Child-Robot Interaction (CRI), 165
- Choreography, 209, 210, 212, 213, 215
- COCO dataset, 361
- Color versus black and white, 183
- Common Cue Hypothesis, 49–51
- Complexity, 372, 382
- Composition, 197, 369
- Compound social cues, 48
- Computer vision, 106, 128, 219
- Content
 - image, 359
- Contextual, 279, 282, 283, 287, 288
- Contrast, 375
- Coronasphere, 205

Cross-domain testing, 235, 243

Crowther, Paul, 221

Culture, 11

D

Dance, 204, 207, 216

Darwin, 4

Data set

- small, 357

dedication, v

Deep feature representations, 235, 240

Deep learning, 223, 231, 232, 242, 249

Deep neural network, 351

Demographic Differences, 331

Depth-of-field, 196

Digital humans, 307

Discrete cosine transform (DCT), 114

E

Education, 334

Effort, 206, 213

Embedding loss, 167

EmoReact, 172

Emotion

- aesthetic, 7

- basic emotions, 7

- componential theories, 10

- continuous dimensions, 9

- emotion models, 3

- recognition, 108, 120, 131, 138

- utilitarian, 7

Emotional movements, 316, 320, 321

Emotion analysis, 294, 295

Emotion expression, 45, 50

Emotion perception, 45, 47–51, 53, 59

Emotion recognition, 314, 316, 318, 324, 326, 327

Emotion Recognition Score (ERS), 171

Emotion residue, 153, 154

Emotion synthesis, 297

Emphasis, 199

Ethics, 92, 93, 97

Ethnicity, 334

Evaluation, 65–67

ExpressionFlow, 129, 131

Eye gaze, 45–48, 73, 74, 81, 83–85

F

Face detector, 124

Face perception, 45–47, 49, 51–54, 147, 149–151, 158

Facial Action Coding System (FACS), 106

Facial Dynamics Map (FDM), 129

Facial expression, 46, 48, 50, 53, 72, 73, 81, 84, 85, 106, 112, 127, 154, 155

Factors used in technical quality prediction, 236

Fear, 46–51, 315, 320–327

Feature importance, 243–246

Feature selection, 246

Feelings, 5

Feminism, 207

Feminist, 91–93, 98, 100

Five-factor model of personality, 294, 295

foreword, vii

Formalism, 222

Frame, 194

Frobenius norm, 112

F1-score, 117

Functional Magnetic Resonance Imaging (fMRI), 76, 81, 83–85

G

Gabor, 114

Gender, 147, 149, 153, 156, 157, 159

Gender differences, 332

Gender-sex, 45, 47–53

Generalization, 25

Generative Adversarial Network (GAN), 298

GloVe, 168

H

Happiness, 49–51, 53, 315, 320–327

High speed, 106, 114

Human-machine interaction, 314, 316

Human-robot interaction (HRI), 204

Human subject study, 333

Hyperparameter, 31

I

Image aesthetics assessment, 231, 234

Image affect modeling, 279–289

Image deblurring, 259

Image denoising, 259

Image inpainting, 260

Image quality and aesthetics, 233, 242, 247

Image quality assessment, 231, 234

Image summarization, 352

Image super-resolution, 260

Image tagging, 106, 108, 123

Income, 334

Infrared reflectography, 353

Interaction effect, 340
Interpersonal stance, 7
Interpretability, 106
Intersectionality, 45, 92
Intersection over Union metric, 363
IoU metric, 363

J

Joint learning, 112

K
Kandinsky, Wassily, 355
Katz, Alex, 362
Kelly, Ellsworth, 358
Kinesphere, 205
Koniocellular pathway, 71, 82
Kubler, George, 220

L

Laban Movement Analysis (LMA), 295, 317, 318, 326
Labor, 91, 92, 95–99
Latent demographics group, 341
Latent dirichlet allocation, 341
Learning
 one-shot, 356
Leveling, 197
Liquid Warping GAN, 301
Local Binary Pattern (LBP), 134
Lost art
 recovery, 355
Low-light enhancement, 261

M

Machine learning, 112, 138, 232, 233, 249, 369
Macroexpression, 128
Magnetoencephalography (MEG), 80, 81, 83–85
Magnocellular pathway, 71, 81, 82
Main effect, 339
Manovich, Lev, 222
Marginalization, 92, 95
Markov chain, 379
Mass, 370
Meaning
 in art, 353
Microexpression, 128
 categorization, 131
 identification, 131
Mirror neurons, 314–316

Modigliani, Amedeo, 362
Mondrian, Piet, 354, 357
Monet, Claude, 358
Mood, 5, 7
Motif, 206, 213–215
Motor expression, 318, 321–324
Movement analysis, 204
Multimedia, 106
Multimodal analyses, 279, 287
Multi-stage regression, 338
Multi-stream, 164, 166

N

Neel, Alice, 362
Nelder-Mead simplex algorithm, 133
Neural network, 36
Neutral faces, 47, 50, 51, 83–85, 150–154, 156–158
No Free Lunch Theorem, 30, 357
Non-physical convention, 354
Non-realistic object, 354
Notation, 204–206

O

OCEAN personality model, 299
On-device learning, 279, 287, 288
Oppression, 91–94, 96
Optical flow, 129
Optimization, 111, 133
Overfitting, 27

P

Parvocellular pathway, 71, 81, 82
Perception, 294, 298
Performance design, 208
Personality, 5, 7
Personality synthesis, 297
Picasso, Pablo, 353, 359, 362
Pixel classification, 360
Point-light display, 61
Pollock, Jackson, 355, 358
Precision, 116
preface, ix
Preference, 5
Primary visual elements, 183
Psychometrics, 249

Q

Quality factor, 248, 250
Quantile regression, 382
Question answering, 352

R

Race, 45, 48–53, 147, 153, 156, 157
 Recall, 117
 Recognition, 106, 113
 Reflection removal, 260
 Regression, 369
 Regularization, 30, 357
 Resistance, 92, 95, 97, 98, 100
 Robotics, 204, 207, 216
 ROC AUC, 173
 Rothko, Mark, 357

S

Saccades, 79, 82, 83
 Sadness, 47, 53, 316, 320–327
 Schemata, 222
 Segmentation, 222, 360
 semantic, 360
 Semantic analysis, 355
 Semantic image segmentation, 352
 Semantic segmentation, 352
 Sexual orientation, 59, 61, 64, 65, 67
 Shannon, Claude, 224
 Shape, 206, 213, 215, 372
 Shared Signal Hypothesis, 47
 Sharpening, 197
 Small data set, 353
 Social categorization, 61, 64
 Social perception, 60–63, 65
 Social Vision, 45
 Somatics, 205, 207, 216
 Spectrogram CNN, 170
 Standpoint, 91, 92
 Statistical anomaly, 220
 Statistical technique, 14
 Stella, Frank, 355
 Stereotypes, 48–52, 54, 59, 65, 149, 153, 156,
 157
 Style, 219, 353
 image, 359
 transfer, 359
 Subjective, 281–285, 287
 Supervised learning, 26
 Support Vector Machine (SVM), 117, 133
 Surrogate art, 361
 Surrogate image, 359

T

Technical and aesthetic quality, 232, 237, 240,
 249, 250
 Technical image quality, 234, 240
 Temporal segment networks, 166
 Ten Item Personality Inventory (TIPI), 299
 Tensor subspace, 136
 Time, 195
 Tonal contrast, 184
 Tonal gradation, 184
 Tone, 374
 Turner, J.M.W., 357

U

Uncanny Valley, 152
 Underfitting, 27
 Unity, 369, 381
 Unsupervised learning, 26

V

Valence, Arousal, and Dominance (VAD), 10,
 171
 Vanitas genre, 355
 Vanitas painting, 353
 Vantage point, 193
 Variance, 29
 Variety, 369, 381
 Virtual agents, 307
 Visual features, 235–237, 241
 Visual sermon, 356
 Visual significance, 378
 Visual stimuli, 332
 Von Jawlensky, Alexej, 362

W

Wilde, Carolyn, 221
 Wollheim, Richard, 222

X

X-radiography, 353