

The Pennsylvania State University
The Graduate School

**TOWARDS AUTOMATED RECOGNITION OF BODILY EXPRESSION OF
EMOTION IN THE WILD**

A Dissertation in
Information Sciences and Technology
by
Yu Luo

© 2021 Yu Luo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2021

The dissertation of Yu Luo was reviewed and approved by the following:

James Z. Wang
Professor of Information Sciences and Technology
Co-Chair of Committee, Dissertation Advisor

Jia Li
Professor of Statistics
Co-Chair of Committee, Dissertation Advisor

Mary Beth Rosson
Professor of Information Sciences and Technology
Associate Dean of Graduate and Undergraduate Studies

Zihan Zhou
Assistant Professor of Information Sciences and Technology

Reginald B. Adams, Jr.
Associate Professor of Psychology

Abstract

Humans are arguably innately prepared to comprehend others’ emotional expressions from subtle body movements. If robots or computers can be empowered with this capability, a number of robotic applications become possible. Automatically recognizing human bodily expression in unconstrained situations, however, is daunting given the incomplete understanding of the relationship between emotional expressions and body movements. The current research, as a multidisciplinary effort among computer and information sciences, psychology, and statistics, proposes a scalable and reliable crowdsourcing approach for collecting in-the-wild perceived emotion data for computers to learn to recognize body languages of humans. To accomplish this task, a large and growing annotated dataset with 9,876 video clips of body movements and 13,239 human characters, named BoLD (Body Language Dataset), has been created. Comprehensive statistical analysis of the dataset revealed many interesting insights. A system to model the emotional expressions based on bodily movements, named ARBEE (Automated Recognition of Bodily Expression of Emotion), has also been developed and evaluated. Our analysis shows the effectiveness of Laban Movement Analysis (LMA) features in characterizing arousal, and our experiments using LMA features further demonstrate computability of bodily expression. We report and compare results of several other baseline methods which were developed for action recognition based on two different modalities, body skeleton and raw image. The dataset and findings presented in this work will likely serve as a launchpad for future discoveries in body language understanding that will enable future robots to interact and collaborate more effectively with humans.

Computationally representing human body movements from images is another aspect towards automated recognition of bodily expression. A fine-grained mesh of human pose and shape provides rich geometric information that enables many applications including bodily expression recognition. Estimating an accurate 3D human mesh from an image captured by a passive sensor is a highly challenging research problem. The mainstream approach, which uses deep learning, requires large-scale human pose/shape annotations in the training process. Currently, those annotations are mostly created from expensive indoor motion capture systems, thus both diversity and quantity are limited. We propose a new method to train a deep human mesh estimation model using a large quantity of unlabeled RGB-D images, which are inexpensive and convenient to collect. Depth information encoded in the data is used in the training process to achieve higher model accuracy. Our method is easy-to-implement and amenable to any other state-of-the-art parametric mesh modeling framework. We empirically demonstrate the effectiveness of this method based on real-world datasets, validating the value of the proposed “learning from depth” approach.

Table of Contents

List of Figures	vii
List of Tables	xi
Acknowledgments	xiii
Chapter 1	
Introduction	1
1.1 Bodily Expression Recognition	1
1.2 Applications of Bodily Expression Recognition	2
1.3 Challenges of Bodily Expression Recognition	3
1.4 Contributions	4
1.5 Structure of Dissertation	5
Chapter 2	
Background: Psychological Perspective of Emotions	7
2.1 History of Psychologist’s Understanding	7
2.2 Two Theoretical Models for Representing Affective States	9
2.3 Elicited Emotions vs. Spontaneous Emotions	10
2.4 Summary	10
Chapter 3	
Background: Computational Modeling of Emotions	12
3.1 Unimodal Emotion Recognition	12
3.1.1 Feature Extraction and Modeling	12
3.1.1.1 Facial Expression	12
3.1.1.2 Bodily Expression	15
3.1.1.3 Speech Expression	18
3.1.1.4 Physiology Indicators	18
3.1.2 Application	19
3.1.3 Comparison	19
3.2 Multimodal Emotion Recognition	19
3.3 Crowdsourced Affect Annotation	21
3.4 Human Pose Representation	22

3.4.1	2D Pose Estimation	22
3.4.2	3D Pose Estimation	24
3.4.3	Human Pose and Shape Reconstruction	25
3.5	Summary	25

Chapter 4

The BoLD Dataset		26
4.1	Dataset Construction	26
4.1.1	Movie Selection and Time Segmentation	27
4.1.2	Pose Estimation and Tracking	28
4.1.3	Emotion Annotation	29
4.1.4	Annotation Quality Control	31
4.1.5	Annotation Aggregation	33
4.2	Dataset Statistics	34
4.2.1	Annotation Distribution and Observations	34
4.2.2	Annotation Quality and Observations	42
4.2.3	Human Performance	44
4.2.4	Demographic Factors	47
4.2.5	Demographic Stereotype	48
4.3	Summary	48

Chapter 5

Bodily Expression Recognition		50
5.1	Learning from Skeleton	50
5.1.1	Laban Movement Analysis	50
5.1.2	Spatial Temporal Graph Convolutional Network	53
5.2	Learning from Pixels	53
5.2.1	Trajectory based Handcrafted Features	54
5.2.2	Deep Activity Features	54
5.3	Results	56
5.3.1	Evaluation Metrics	56
5.3.2	LMA Feature Significance Test	56
5.3.3	Model Performance	57
5.3.4	Ablation Study	59
5.3.5	ARBEE: Automated Recognition of Bodily Expression of Emotion	60
5.4	Summary	62

Chapter 6

Learning from Depth: Image-based 3D Human Mesh Reconstruction using Unlabeled RGB-D Data		63
6.1	Introduction	63
6.2	Related Work	65
6.3	Technical Approach	67
6.3.1	The SMPL Model	67

6.3.2	Depth-aware Ground Truth Generation	68
6.3.3	Learning Solution	69
6.4	Experiments	70
6.4.1	Training Datasets	70
6.4.2	Evaluation Datasets and Metrics	73
6.4.3	Experiment Setup	74
6.4.4	Ablation Study	76
6.4.5	Quantitative Results	77
6.4.6	Qualitative Results	77
6.5	Summary	78
Chapter 7		
	Conclusions and Future Work	79
7.1	Summary	79
7.2	Future Work	79
Appendix		
	IVAR: Image Variation Retrieval via Feature Learning	81
A.1	Introduction	82
A.2	Related Work	83
A.3	Analysis of Image Transformations	84
	A.3.1 Simulated Image Transformations	86
	A.3.2 Analysis of Image Transformations	87
A.4	The Approach	89
A.5	Experiments	91
	A.5.1 Datasets	91
	A.5.2 Evaluation Metric	92
	A.5.3 Implementation Details	93
	A.5.4 Experimental Results	93
	A.5.5 Discussion	95
A.6	Conclusions	96
Bibliography		97

List of Figures

1.1	Examples of possible scenarios where computerized bodily expression recognition can be useful. From left to right: psychological clinic assistance, public safety and law enforcement, and social robot or social media.	2
2.1	PAD-model. Image from Figure 1 of [1]	8
3.1	Part of face action coding system. Image from Table 19.1 in [2]	13
3.2	Motion capture system. Image from http://movies-in-theaters.net/mocap-software.html	16
3.3	Three different human pose representation: 2D pose, 3D pose and human pose mesh.	23
3.4	2D models of human shape. From the left: Cardboard People (in two view-points), Pictorial Structures (PS), Contour People (CP) and Deformable Structures (DS). Image from Figure 2 of [3].	24
4.1	Overview of our data collection pipeline. The process involves crawling movies, segmenting them into clips, estimating the poses, and emotion annotation.	26
4.2	Selected movie titles.	28
4.3	A frame in a video clip, with different characters numbered with an ID (<i>e.g.</i> , 0 and 1 at the bottom left corner of red bounding boxes) and the body and/or facial landmarks detected (indicated with the stick figure).	29
4.4	The web-based crowdsourcing data collection process. Screenshots of the four steps are shown. For each video clip, participants are directed to go through a sequence of screens with questions step-by-step.	30

4.5	Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion. To be continued on the next page.	35
4.5	(Continued from the previous page.) Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances (27 and 28) that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion.	36
4.5	(Continued from the previous page.) Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances (27 and 28) that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion.	37
4.5	(Continued from the previous page.) Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances (27 and 28) that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion.	38
4.5	(Continued from the previous page.) Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances (27 and 28) that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion.	39
4.6	Distributions of the 26 different categorical emotions.	40

4.7	Distributions of the three dimensional emotion ratings: valence, arousal, and dominance.	41
4.8	Demographics of characters in our dataset.	41
4.9	Correlations between pairs of categorical or dimensional emotions, calculated based on the BoLD dataset.	42
4.10	Reliability score distribution among low-performance participants (failure) and non low-performance participants (pass).	45
4.11	Human regression performance on dimensional emotions. X-axis: participant population percentile. Y-axis: $F1$, R^2 and MSE score. Tables inside each plot in the second row summarize top 30%, 20%, 10%, and 5% participant regression scores.	46
5.1	Illustration of the human skeleton. Both red lines and black lines are considered limbs in our context.	52
5.2	Kernel density estimation plots on selected LMA features that have high correlation with arousal.	55
5.3	Classification performance (AP: average precision on the top left, RA: ROC AUC on the top right) and regression performance (R^2 on the bottom) of different methods on each categorical and dimensional emotion.	58
6.1	Our method optimizes the human oracle from an image-based initialization to one being also consistent with additional depth data. The resulting optimized human model parameters are then used as supervision for a machine learning system that only takes a 2D image as input and generates human pose and shape, <i>i.e.</i> , parameters of the SMPL parametric model [4].	64
6.2	Distribution of absolute error of rendered depth under different camera perspectives. Each bar in the figure corresponds to a a specific camera perspective. SPIN Pseudo GT is more sensitive to camera perspectives compared with depth-aware GT and the prediction of our fine-tuned model.	72

6.3	Qualitative results of SPIN GT and ours Depth-aware GT. The upper panel examples are from HDE dataset and the lower panel ones are from NTU- RGBD test set. For each panel, the first row is the original RGB image. The second and third row are the front view of the human mesh from SPIN GT and Depth-aware GT. The forth and fifth row are the side view of the human mesh from SPIN GT and Depth-aware GT.	75
7.1	Layers of abstraction for bodily expressed emotions.	80
A.1	Examples of two common image manipulation processes – compositing and retouching. The right three images are manipulated from the first image on the left. Specifically, the second and the third image are edited via compositing, <i>i.e.</i> , inserting the man and the woman into the image. The fourth and the fifth image are retouched with different artistic filters.	81
A.2	An instance in our training set. The removing example is generated by remov- ing the boy out of the image. The inserting example is generated by inserting the car. The filters example is generated with a Gaussian blur filter. The WCT and photoWCT examples are generated by two different style transfer algorithms.	85
A.3	Distance distribution of baseline model. Histograms of intra-instance dis- tance, inter-instance distance, and inter-class distance on the off-the-shelf CNN feature space.	86
A.4	Intra-instance distance distribution over different transformations. The blue histogram is the distribution of intra-instance distance between pairs in the title, <i>e.g.</i> , raw and remove for the first figure. The green histogram is the distribution of inter-instance distance. The red histogram is the distribution of inter-class distance. The farer intra-instance distance is from the other two, the more possible that a model retrieves desired image variants.	87
A.5	Objective illustration. Squares with the same color are from the same instance. Our goal is to increase inter-instance distance d_- and decrease intra-instance distance d_+	90
A.6	Example retrieval results on the Places-IVAR-Test dataset. The first column is the query image. For each two rows, the first row is retrieved using the baseline model and the second row is retrieved using our model. Similarity decreases from the left to the right.	94
A.7	Distance distribution of our model. Histograms of intra-instance distance, inter-instance distance, and inter-class distance on learned feature space.	95

List of Tables

3.1	Comparison on different emotion recognition modalities.	20
4.1	Agreement among participants on categorical emotions and characters' demographic information.	43
5.1	Laban Movement Analysis (LMA) features. (f_i : categories; m : number of measurements; dist.: distance; accel.: acceleration)	51
5.2	Dimensional emotion regression and categorical emotion classification performance on the test set. mR^2 = mean of R^2 over dimensional emotions, mAP(%)= average precision / area under precision recall curve (PR AUC) over categorical emotions, mRA(%) = mean of area under ROC curve (ROC AUC) over categorical emotions, and ERS = emotion recognition score. Baseline methods: ST-GCN [5], TF [6], TS-ResNet101 [7], I3D [8], and TSN [9]. . .	57
5.3	Ablation study on the effect of pretrained models.	60
5.4	Ablation study on the effect of face.	60
5.5	Ensembled results.	60
5.6	Retrieval results of our deep model. P@K(%) = precision at K, R-P(%)=R-Precision.	61

6.1	Various datasets used and their availability of different supervision signals: 2D pose, 3D pose, and SMPL parameters. ✓ means the supervision signal is manually annotated or directly recorded from the sensor. ✓* indicates the supervision signal being estimated from an oracle approach. ✗ indicates the supervision signal being unavailable. Note that SMPL parameters cannot be either manually annotated or directly recorded due to the nature of the SMPL parametric model. We refer readers to the corresponding references on how SMPL parameters were estimated. We remark that MoSh [10] is accurate enough to be treated as “ground truth” due to the usage of dense MoCap markers. Newly added datasets in our work, <i>i.e.</i> , NTU RGB+D and HDE dataset, have their 2D pose estimated from OpenPose [11].	71
6.2	Experiment setup comparison. An oracle SMPL parameters initialization and 2D pose can be easily available from any pre-trained models, while RGB and depth images can be directly recorded from various mobile devices with a depth sensor.	73
6.3	Ablation study: Evaluation on the rendered depth for NTU-RGBD test set. . .	74
6.4	Evaluation on the reconstructed 3D pose over different datasets. Both MPJPE and PA-MPJPE has the unit of millimeter.	74
6.5	Evaluation on the reprojected pose and shape over LSP test set. Reconstructed human mesh is reprojected to the image plane to evaluate foreground-background and six-part segmentation.	76
6.6	Evaluation on the rendered depth over HDE dataset. Our approach outperforms both the pre-trained model [12] and the models fine-tuned under the same semi-supervised setup while different pseudo ground truth.	77
A.1	Evaluation dataset statistics.	91
A.2	Results on evaluation dataset. Image variation retrieval results in terms of mAP are reported on three evaluation datasets. On Places-IVAR-Test, top 1 classification accuracy is also reported in the bracket.	92

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Prof. James Z. Wang and my co-advisor Prof. Jia Li, for their invaluable advice, encouragement and support. I am extremely fortunate to have them as my mentors along this marvelous journey. Thanks to them, I had the chance to learn novel knowledge through challenging research problems. With their encouragement, I have developed the courage to insist on the highest standards for my work.

Next I would like to thank Dr. Mary Beth Rosson, Dr. Zihan Zhou and Dr. Reginald B. Adams for serving as my dissertation committee members. Dr. Reginald B. Adams and Dr. Michelle G. Newman gave insightful discussion and invaluable guidance from a psychological perspective, and I am deeply grateful to them.

My gratitude also extends to my friends and colleagues within and outside Penn State. I have been extremely fortunate to work with excellent collaborators and lab mates, and I am thankful for each one: Jianbo Ye, Xin Lu, Hanjoo Kim, Baris Kandemir, Xinye Zheng, Yukun Chen, Mohanmmad Kamani, Farshid Farhat, Zhuomin Zhang and Dolzodmaa Davaasuren. I also owe special appreciation to my friends: Jiasheng Zhang, Chen Liang, Xiao Yang, Mo Yu, Haoti Zhong, Hongjian Wang and Fei Wu. I will always cherish the days we shared together at the lovely small town of State College.

I would like to express special thanks to the thousands of Amazon Mechanical Turk independent contractors for their time and dedication in providing invaluable emotion ground truth labels for the video collection.

I am also grateful to the College of Information Sciences and Technology, the US National Science Foundation and Amazon.com, Inc. for financially supporting my Ph.D. study. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

Finally, I am most grateful to my parents for their continued and unconditional support and encouragement. Thanks to them, I had the faith and courage to go through ups and downs in my life.

Dedication

To my parents, for their unconditional support.

Chapter 1 | Introduction

1.1 Bodily Expression Recognition

Many future robotic applications, including personal assistant robots, social robots, and police robots demand close collaboration with and comprehensive understanding of the humans around them. Current robotic technologies for understanding human behaviors beyond their basic activities, however, are limited. Body movements and postures encode rich information about a person's status, including their awareness, intention, and emotional state [13]. Even at a young age, humans can “read” another's body language, decoding movements and facial expressions as emotional keys. *How can a computer program be trained to recognize human emotional expressions from body movements?* This question drives our current research effort.

Previous research on computerized body movement analysis has largely focused on recognizing human activities (*e.g.*, the person is running). Yet, a person's emotional state is another important characteristic that is often conveyed through body movements. Recent studies in psychology have suggested that movement and postural behavior are useful features for identifying human emotions [14–17]. For instance, researchers found that human participants of a study could not correctly identify facial expressions associated with winning or losing a point in a professional tennis game when facial images were presented alone, whereas they were able to correctly identify this distinction with images of just the body or images that included both the body and the face [17]. More interestingly, when the face part of an image was paired with the body and edited to an opposite situation face (*e.g.*, winning face paired with losing body), people still used the body to identify the outcome. A valuable insight from this psychology study is that the human body may be more diagnostic than the face in terms of emotion recognition. In our work, *bodily expression* is defined as human affect expressed by body movements and/or postures.

Our earlier work studied the computability of evoked emotions [18–20] from visual stimuli



Figure 1.1: Examples of possible scenarios where computerized bodily expression recognition can be useful. From left to right: psychological clinic assistance, public safety and law enforcement, and social robot or social media.

using computer vision and machine learning. In this work, we investigate whether bodily expressions are computable. In particular, we explore whether modern computer vision techniques can match the cognitive ability of typical humans in recognizing bodily expressions in the wild, *i.e.*, from real-world unconstrained situations.

1.2 Applications of Bodily Expression Recognition

Computerized bodily expression recognition capabilities have the potential to enable a large number of innovative applications including information management and retrieval, public safety, patient care, and social media [21]. For instance, such systems can be deployed in public areas such as airports, metro or bus stations, or stadiums to help police identify potential threats. Better results might be obtained in a population with a high rate of emotional instability. A psychology clinic, for example, may install such systems to help assess and evaluate disorders, including anxiety and depression, either to predict danger to self and others from patients, or to track the progress of patients over time. Similarly, police may use such technology to help assess the identity of suspected criminals in naturalistic settings and/or their emotions and deceptive motives during an interrogation. Well-trained and experienced detectives and interrogators rely on a combination of body language, facial expressions, eye contact, speech patterns, and voices to differentiate a liar from a truthful person. An effective assistive technology based on emotional understanding could substantially reduce the stress of police officers as they carry out their work. Improving the bodily expression recognition of assistive robots will enrich human-computer interactions. Future assistive robots can better assist those who may suffer emotional stress or mental illness, *e.g.*, assistive robots may detect

early warning signals of manic episodes. In social media, recent popular social applications such as Snapchat and Instagram allow users to upload short clips of self-recorded and edited videos. A crucial analysis from an advertising perspective is to better identify the intention of a specific uploading act by understanding the emotional status of a person in the video. For example, a user who wants to share the memory of traveling with his family would more likely upload a video capturing the best interaction moment filled with joy and happiness. Such analysis helps companies to better personalize the services or to provide advertisement more effectively for their users, *e.g.*, through showing travel-related products or services as opposed to business-related ones.

1.3 Challenges of Bodily Expression Recognition

Automatic bodily expression recognition as a research problem is highly **challenging** for three primary reasons. First, it is difficult to collect a bodily expression dataset with high quality annotations. The understanding and perception of emotions from concrete observations is often subject to context, interpretation, ethnicity and culture. There is often no gold standard label for emotions, especially for bodily expressions. In facial analysis, the expression could be encoded with movements of individual muscles, a.k.a., Action Units (AU) in facial action coding system (FACS) [22]. However, psychologists have not developed an analogous notation system that directly encodes correspondence between bodily expression and body movements. This lack of such empirical guidance leaves even professionals without complete agreement about annotating bodily expressions. To date, research on bodily expression is limited to acted and constrained lab-setting video data [23–26], which are usually of small size due to lengthy human subject study regulations. Second, bodily expression is subtle and composite. According to [1], body movements have three categories, functional movements (*e.g.* walking), artistic movements(*e.g.* dancing), and communicative movements (*e.g.* gesturing while talking). In a real-world setting, bodily expression can be strongly coupled with functional movements. For example, people may represent different emotional states in the same functional movement, *e.g.* walking. Third, an articulated pose has many degrees of freedom. Working with real-world video data poses additional technical challenges such as the high level of heterogeneity in people’s behaviors, the highly cluttered background, and the often substantial differences in scale, camera perspective, and pose of the person in the frame.

1.4 Contributions

In this work, we investigate the feasibility of crowdsourcing bodily expression data collection and study the computability of bodily expression using the collected data and explore the intuitive idea of “learning from depth” for better human mesh reconstruction. We summarize the primary **contributions** as follows.

- We propose a scalable and reliable crowdsourcing pipeline for collecting in-the-wild perceived emotion data. With this pipeline, we collected a large dataset with 9,876 clips that have body movements and over 13,239 human characters. We named the dataset the **BoLD (Body Language Dataset)**. Each short video clip in BoLD has been annotated for emotional expressions as perceived by the viewers. To our knowledge, BoLD is the first large-scale video dataset for bodily emotion in the wild.
- We conducted comprehensive agreement analysis on the crowdsourced annotations. The results demonstrate the validity of the proposed data collection pipeline. We also evaluated human performance on emotion recognition on a large and highly diverse population. Interesting insights have been found in these analyses.
- We investigated Laban Movement Analysis (LMA) features and action recognition-based methods using the BoLD dataset. From our experiments, hand acceleration shows strong correlation with one particular dimension of emotion — arousal, a result that is intuitive. We further show that existing action recognition-based models can yield promising results. Specifically, deep models achieve remarkable performance on emotion recognition tasks.
- We developed a method easy for incorporation into the existing human mesh reconstruction framework to gain “depth” knowledge from RGB-D training data. We evaluated the model accuracy from the depth perspective in addition to other conventional metrics to demonstrate the complementary benefits of semi-supervised learning and “learning from depth”.

In our work, we approach the bodily expression recognition problem with the focus of addressing the first and third challenge mentioned earlier. Using our proposed data collection pipeline, we have collected high quality affect annotation. With our proposed “learning-from-depth” pipeline, we are able to train a better model for human mesh reconstruction with finer-grained details. We believe the third challenge can be addressed when more and more RGB-D data are collected and fitted into our “learning-from-depth” pipeline. To properly address the second challenge, regarding the subtle and composite nature of bodily expression,

requires breakthroughs in computational psychology. Below, we detail some of the remaining technical difficulties on the bodily expression recognition problem that the computer vision community can potentially address.

Despite significant progress recently in 2D/3D pose estimation [27, 28], these techniques are limited compared with Motion Capture (MoCap) systems, which rely on placing active or passive optical markers on the subject’s body to detect motion, because of two issues. First, these vision-based estimation methods are noisy in terms of the jitter errors [29]. While high accuracy has been reported on pose estimation benchmarks, the criteria used in the benchmarks are not designed for our application which demands substantially higher precision of landmark locations. Consequently, the errors in the results generated through those methods propagate in our pipeline, as pose estimation is a first-step in analyzing the relationship between motion and emotion.

Second, vision-based methods (*e.g.*, [28]) usually address whole-body poses, which have no missing landmarks, and only produce relative coordinates of the landmarks from the pose (*e.g.*, with respect to the barycenter of the human skeleton) instead of the actual coordinates in the physical environment. In-the-wild videos, however, often contain upper-body or partially-occluded poses. Further, the interaction between human and the environment, such as a lift of the person’s barycenter or when the person is pacing between two positions, is often critical for bodily expression recognition. Additional modeling on the environment together with that for the human would be useful in understanding body movement.

In addition to these difficulties faced by the computer vision community broadly, the computation psychology community also needs some breakthroughs. For instance, state-of-the-art end-to-end action recognition methods developed in the computer vision community offer insufficient interpretability of bodily expression. While the LMA features that we have developed in this work has better interpretability than the action recognition based methods, to completely address the problem of body language interpretation, we believe it will be important to have comprehensive motion protocols defined or learned, as a counterpart of FACS for bodily expression.

1.5 Structure of Dissertation

The rest of this dissertation is structured as follows. Chapter 2 and Chapter 3 review related work in the literature from psychological and computational perspectives respectively. The data collection pipeline and statistics of the BoLD dataset are introduced in Chapter 4. We describe our modeling processes on BoLD and demonstrate findings in Chapter 5. We introduce

our method to reconstruct expressive human pose and shape in Chapter 6 and conclude in Chapter 7.

Chapter 2 |

Background: Psychological Perspective of Emotions

2.1 History of Psychologist's Understanding

Studies on affective states and their physical representations could be dated back to 1870s [30]. In this book, emotion/mental states were linked to their biological representations for the first time. Darwin started with special movements and found their strong correlation to certain kinds of emotion. Thereafter, following the ideas of Darwin, James [31] demonstrated that bodily changes are not necessarily the consequence of emotions. In other words, postures and gestures are actually part of emotions. Darwin and James's work establish the area of emotion study and define the emotion by including bodily representation as part of the emotion. Also, their research methods (questionnaire and experiments methods) provide guidance for the future research.

In 1970s, Ekman *et al.* published several seminal papers about facial expressions and emotion study [22, 32–34]. In [34], Ekman found out universal rules for facial expressions. Based on the findings, Facial Action Coding System (FACS) was proposed to describe different facial expressions by observing the status of facial muscles [22]. From these studies, Ekman argued that anger, happiness, sadness, surprise, disgust, and fear are the most basic emotions [32]. To conclude, Ekman's work defines emotions further and the conclusions provide the computability of emotion for computer scientists since the subjective activity can be represented by finite objective classes. At the same time, as shown in Figure 2.1, Mehrabian proposed another PAD-model [35] that measures affective states in three dimensions, valence (the extent of pleasure), arousal (the intensity of physical activity) and dominance (the extent of feeling controlled) was proposed. These two models are widely used by computer scientists and further discussion about these two models can be found in Section 2.2.

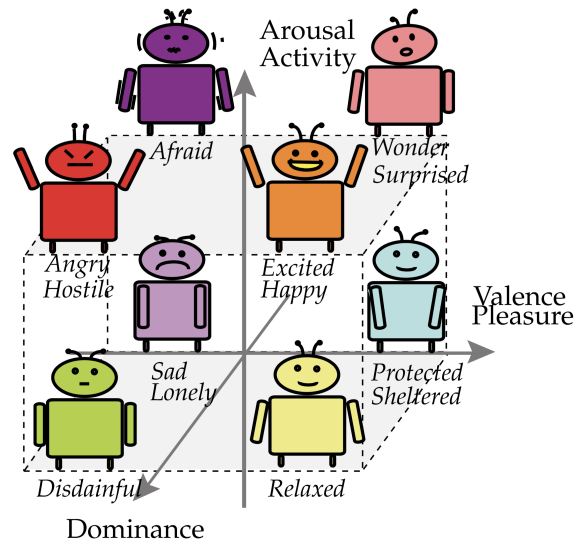


Figure 2.1: PAD-model. Image from Figure 1 of [1]

Right after Ekman's facial expression work, psychologists began to find out how much body movements [14] and vocal cues [36] could contribute to emotions. Similarly, psychologists use movement notation system to describe movements like FACS in facial expression. In [14], Wallbott used a simple notation system to encode actors' movements while they are performing certain emotions under some scenarios. Through statistical analysis, Wallbott concluded that some movements showed a strong correlation with certain emotions. To date, there are more and more evidence [26, 37] showing the importance of bodily expression. The results in [14], for the first time, indicate the existence of some emotion-specific movements and postures in a statistical analysis. Vocal cues, according to [36], have the similar characteristic to facial expression since they represent affective states in both valence and arousal dimension.

Despite the contribution of bodily expression to affective states, there are debates about how exactly bodily expression contribute to affective states. Previously, it was assumed that bodily expression can only represent the intensity of emotions [38]. However, recent studies [17, 39] indicate that body cues could outperform facial expressions in discriminating positive and negative emotions.

Currently, bodily expressions are still understudied compared with facial expression. In [17], it seems that facial expressions contribute less in terms of intense positive and negative emotions compared with bodily expressions. In [39], the results suggest anger is easier to recognize from bodily expression. Psychologists have begun to investigate influences of affective states and body postures/gestures on each other [40]. In this book [40], Cuddy argues that people could

grow positive attitude, increase their confidence and improve their affective states through performing positive and dominant postures or gestures. There are also studies on other factors that may influence bodily expressions, such as culture [41] and gender [42].

2.2 Two Theoretical Models for Representing Affective States

Existing automated bodily expression recognition studies mostly build on two theoretical models for representing affective states, the *categorical* and the *dimensional* models. The categorical model represents affective states into several emotion categories. In [32,43], Ekman *et al.* proposed six basic emotions, *i.e.*, anger, happiness, sadness, surprise, disgust, and fear. However, as suggested by [44] and [1], bodily expression is not limited to basic emotions. When we restricted interpretations to only basic emotions at a preliminary data collection pilot study, the participants provided feedback that they often found none of the basic emotions as suitable for the given video sample. A dimensional model of affective states is the PAD model by [35], which describes an emotion in three dimensions, pleasure (valence), arousal, and dominance. In the PAD model, valence characterizes the positivity versus negativity of an emotion, while arousal characterizes the level of activation and energy of an emotion, and dominance characterizes the extent of controlling others or surroundings. As summarized in [1,45], most bodily expression-related studies focus on either a small set of categorical emotions or two dimensions of valence and arousal in the PAD model. In our work, we adopt both measurements in order to acquire complementary emotion annotations.

There have been discussions and debates on different view of emotions [46]. Some theorists believe emotion is triggered by the brain's appraisal of a stimulus. They are distinctive and subtle emotions are combination of different basic emotions. Other theorists believe emotion is constructed from a few emotional ingredients/dimensions. To date, there is little physiology study in favor of any of the theory. From a practitioner's perspective, we should choose the representation based on the real application. For example, we can focus on fatigue/boredom in an application for driver fatigue detection; a dimensional representation could be adopted to monitor people's valence in an application of tracking elderly mental health at home.

2.3 Elicited Emotions vs. Spontaneous Emotions

Based on how emotion is generated, emotions can be categorized into acted or elicited emotions, and spontaneous emotions. Acted emotion refers to actors' performing a certain emotion under given contexts or scenarios. Early work was mostly built on acted emotions [14,23,25,26]. [14] analyzes videos recorded on recruited actors and established bodily emotions as an important modality of emotion recognition. In [47], a human subject's emotion is elicited via interaction with computer avatar of its operator. [19] crowdsourced emotion responses with image stimuli. Recently, natural or authentic emotions have generated more interest in the research community. In [48], body movements are recorded while human subjects play body movement-based video games.

2.4 Summary

There are two typical experiment paradigms for facial expression and bodily expression study in psychology. One is asking subjects to label different movements with emotions and finding out emotion-specific characteristics [34], or ask subjects to act specific emotions directly. The other is encoding the physical movements into certain formats and analyzing statistical significance between the formatted data and ground truth/emotions [14]. Both paradigms encode expressions into more abstract features so that statistical significance could be computed. The encoding systems are FACS [22] for facial expressions and movement notation system [49] for bodily expressions. With more sophisticated systems, descriptions could be more accurate and encoding process would be more arduous and vice versa. Based on the specific research question and characteristic of the selected notation system, experiments are designed to corroborate the hypothesis.

Unlike FACS for facial expression, however, there are many different movement notation systems designed for bodily expression. There is a comprehensive discussion on those systems in [1]. The main issue for those movement notation systems would be that coder's perceptual inference is still needed. The reliance on perceptual inference makes the encoding process hard to automatize and the data collection/processing would be extremely arduous subsequently. This is one of the reasons why **available bodily expression datasets are of small size**. Furthermore, coders may have different standards for the inference and the results would be less convincing consequently. Another weakness of current studies would be that most of the datasets collected are acted elicited emotions in laboratory settings.

To conclude, **studies in psychology has confirmed bodily expressions' influence on**

affective states. However, the way it influences and how much it influences are still unclear. Psychologists have tried to extract movement features manually and analysis on lab-setting dataset has shown correlation of affective states and features like dynamic motion, hands configuration and head orientation. These findings have provided insights for designing automatic emotion recognition systems. Firstly, the collected data should be **unbiased and balanced** by considering other factors like culture and gender. Secondly, the systems should be able to **extract features** like movement notation systems.

Chapter 3 |

Background: Computational Modeling of Emotions

3.1 Unimodal Emotion Recognition

Humans perceive and understand emotions from multiple modalities, such as face, body language, touch, eye contact, and vocal cues. There are multiple sensors available for capturing emotion-related information, such as cameras, microphones, motion capture systems, and physiology sensors. For automatic systems, signal processing and image processing would be used firstly for features extraction. In this process, both spatial and temporal information should be retained so that machine learning algorithms could do classification. In designing of features for recognition, domain knowledge is usually used. Besides, there are automatic feature extraction approaches like deep learning.

In subsection 3.1.1, detailed feature extraction and modeling of each modality (facial expression, bodily expression, speech expression, text expression, and other physiology indicators) would be demonstrated. In subsection 3.1.2, several real-world applications from emotion recognition will be introduced. In subsection 3.1.3, advantages and challenges of emotion recognition in each modality will be discussed and summarized.

3.1.1 Feature Extraction and Modeling

3.1.1.1 Facial Expression

Facial expression is an important modality in emotion recognition and automated facial expression recognition is more successful compared with other modalities. The main reasons for this success are two-fold. First, the discovery of FACS made facial expression less subjective. Many

Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
					
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
					
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 3.1: Part of face action coding system. Image from Table 19.1 in [2]

recent works on facial expression recognition focus on Action Unit detection, *e.g.*, [50, 51]. Second, the face has fewer degrees of freedom compared with the whole body [25]. To address the comparatively broader freedom of bodily movement, [1] suggest the use of a movement notation system may help identify bodily expression. Other research has considered microexpressions, *e.g.*, [52], suggesting additional nuances in facial expressions. To our knowledge, no vision-based study or dataset on complete measurement of natural bodily emotions exists.

Psychologists firstly found consistency in facial expression recognition over different cultures [34]. In the later work [22], Ekman proposed facial action coding system (FACS) for describing movements on a face. By adopting FACS, basic emotion categories could be distinguished with strong patterns. Part of FACS is shown in Figure 3.1. As described in [53], automatic facial expression recognition could be divided into three tasks: face detection, facial expression information extraction, and facial expression classification.

For face detection, Viola *et al.* adopt adaptive boosting (AdaBoost) with Harr feature as input for weak classifier [54]. Specifically, two-rectangle feature, three-rectangle feature, and four-rectangle features are used for capturing **texture** of images. The approach achieves real-time detection by combining classifiers in a 'cascade' which reject non-face region at an early stage. The algorithm in [54] and its variations have been widely used nowadays since its robustness and efficiency. However, the performances could still suffer from occlusion, head orientation, and lightning conditions.

For facial expression information extraction, **geometric features and appearance features** are usually considered to capture the information as FACS could achieve. According to [2], geometric features usually capture shape and location of facial components, whereas appearance features extract skin textures and local information. In [55], geometric features are extracted by fitting a 3D face mesh model to the detected face region. Several landmarks on face are selected as geometric features and tracked in videos through optical flow. Appearance features are extracted from certain face regions based on domain knowledge, such as corners of mouth and eyes. Gabor filters are used for capturing local changes and textures in these regions. In [55], Wen *et al.* improve the appearance features by applying Gabor filters on ratio-images, which makes the feature insensitive to face albedo. Compared with using geometric features only, accuracy is increased significantly (about $\sim 20\%$ in average) after using both geometric features and appearance features. Apart from using handcrafted features like Gabor filters, deep learning could also be used for feature extraction and it is actually famous for efficient automatic feature extraction. However, adopting deep learning approaches usually requires a large annotated dataset, which is not available in facial expression recognition research. Several recent studies on facial expression recognition with convolutional neural network (CNN) [56, 57] have been explored by leveraging existing limited data for training CNN and conducting data augmentation.

As the last step of facial expression recognition, models for facial expression classification are usually frame-based or sequential-based [2]. For frame-based models, classifiers are built with features extracted from the current frame as input. Typical methods to achieve that are support vector machine (SVM), artificial neural network (ANN), K nearest neighbor (KNN) and so on. For sequential-based models, hidden Markov model (HMM) is usually used due to its success in temporal pattern recognition.

To conclude, a facial expression recognition system mainly extracts facial landmark configuration and facial texture at some critical regions for facial expression recognition through handcrafted designs of features (Gabor filters) or supervised feature extraction (CNN). Although facial expression recognition has achieved solid progress and techniques are robust

enough for commercial use¹, it may still be cumbersome for real-world use. As suggested in [55], accuracy relies on appearance features partly. With low resolution, facial expression recognition could be unreliable. Also, there are many circumstances where faces are occluded or front faces are not visible at all.

3.1.1.2 Bodily Expression

Automatic modeling of bodily expression (AMBE) typically requires three steps: human detection, pose estimation and tracking, and representation learning. In such a pipeline, human(s) are detected frame-by-frame in a video and their body landmarks are extracted by a pose estimator. Subsequently, if multiple people appear in the scene, the poses of the same person are associated along all frames [58]. With each person’s pose identified and associated across frames, an appropriate feature representation of each person is extracted.

Based on the way data is collected, we divide AMBE methods into video-based and non-video-based. For video-based methods, data are collected from a camera, in the form of color videos. In [59, 60], videos are collected in a lab setting with a pure-colored background and a fixed-perspective camera. They could detect and track hands and other landmarks with simple thresholding and grouping of pixels. [59] additionally defined motion protocols, such as whether the hand is facing up, and combined them with landmark displacement as features. [60] used the positions of shoulders in the image frame, facial expression, and audio features as the input of a neural network. Our data, however, is not collected under such controlled settings, thus has variations in viewpoint, lighting condition, and scale.

For non-video-based methods, locations of body markers are inferred by the MoCap system [24, 25, 48, 61]. The first two steps, *i.e.*, human detection, and pose estimation and tracking, are solved directly by the MoCap system. Geometric features, such as velocity, acceleration, and orientation of body landmarks, as well as motion protocols can then be conveniently developed and used to build predictive models [24, 48, 61]. For a more comprehensive survey of automatic modeling of bodily expression, readers are referred to the three surveys [1, 45, 62].

Related to AMBE, human behaviour understanding (a.k.a. action recognition) has attracted a lot of attention. The emergence of large-scale annotated video datasets [63–65] and advances in deep learning [66] have accelerated the development in action recognition. To our knowledge, two-stream ConvNets-based models have been leading on this task [7–9]. The approach uses two networks with an image input stream and an optical flow input stream to characterize appearance and motion, respectively. Each stream of ConvNet learns human-action-related features in an end-to-end fashion. Recently, some researchers have attempted to utilize human

¹<https://www.microsoft.com/cognitive-services/en-us/emotion-api>

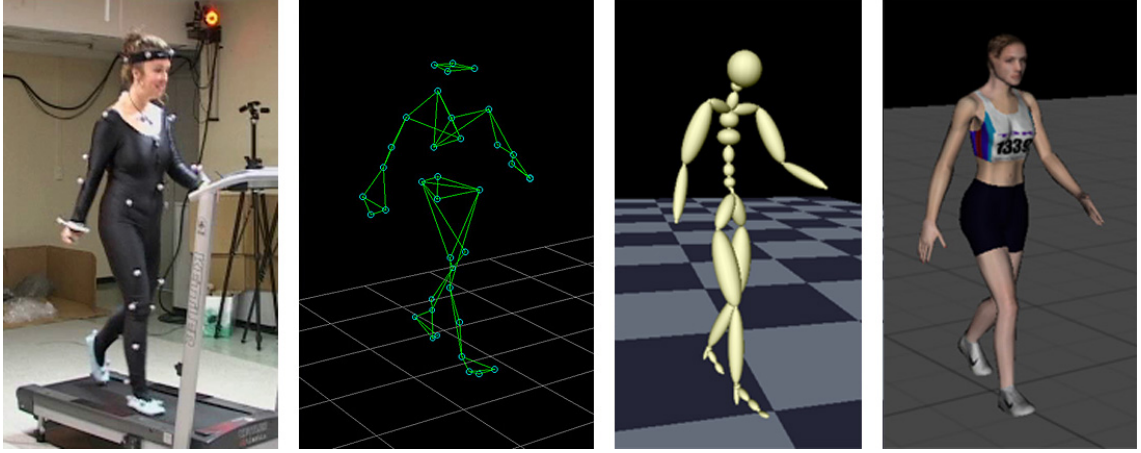


Figure 3.2: Motion capture system. Image from <http://movies-in-theaters.net/mocap-software.html>

pose information. [5], for example, modeled human skeleton sequences using a spatiotemporal graph convolutional network. [67] leveraged pose information using a multitask-learning approach. In our work, we extract LMA features based on skeletons and use them to build predictive models.

Related work can be categorized based on raw data types, namely MoCap data or image/video data. For lab-setting studies such as [24, 48, 61], collecting motion capture data is usually feasible. [23] collected a dataset with upper body movement video recorded in a studio. Other work [23, 25, 47] used image/video data capturing the frontal view of the poses. As shown in Figure 3.2, joints coordinates are recorded by putting marks on each joint. Typically, these kinds of equipment² are for commercial use and comes with calibration software. Therefore, data collected from this sensor are usually normalized joints coordinates and need no further preprocessing. Existing datasets collected from MoCap are usually with acted emotions and of small size [48]. In [24], MoCap data are used to study cross-cultural effect in bodily expression. In this work, Kleinsmith *et al.* use coordinates of joints and velocity as features. Dimension reduction methods like principal component analysis are usually needed for efficient classification. Compared with MoCap data, videos from normal cameras are of large size and could contain spontaneous/natural emotions. Hence, videos is another important data source for bodily expression recognition. Similarly, bodily expression recognition could also be decomposed into three tasks: human body detection, bodily expression information extraction, and bodily expression classification.

Traditionally, human body detection is solved by handcrafted feature Histograms of Ori-

²<https://www.vicon.com/>

ented Gradient (HOG) [68]. Essentially, HOG is a kind of appearance features since it captures the local gradient information of an image. Normalization of cells over larger blocks also makes this feature invariant to illumination changes and shadowing. To date, HOG is still one of the most robust features in computer vision. Its performance on pedestrian detection is rather satisfactory.

For bodily expression information extraction, it is often referred as pose estimation in computer vision community. The prime goal for pose estimation is to infer **human body movements or postures** through locating joints' position like what MoCap could achieve. Like FACS in face research, there are also several different movement notation systems for body, such as Birdwhistell [49]. However, considering that the degree of freedom of body postures/gestures is much larger than that of facial expression, current movement notation system is not as accurate and simple as FACS. Detailed discussion on the difference of motion notation systems is referred to [1]. Previously, pose estimation was explored by following deformable part model (DPM) [69]. In fact, DPM was originally proposed for object detection. In DPM, dense HOG feature maps on each level of image pyramid are taken as input. Subsequently, the features are used to compute the score which reflects the extent of match of two part with the cost of difference from root part. Since DPM explicitly contains possible degrees of freedom in the model, it is suitable for further pose estimation. In [70], the accuracy of pose estimation based on DPM is achieved at around 40%. Obviously, the results are still too weak to be used as features for bodily expression recognition. Recently, researchers have designed many different CNN architectures with consideration of other factors, like optical flow, and the accuracy is approximate to 80% [71, 72]. To date, pose estimation is still a challenging task in computer vision community. However, it may not be necessary to extract exact joints' coordinates for bodily expression recognition. In [73], Kwak *et al.* learn a pose representation through CNN. In this way, similar poses are encoded in the new space with a short distance.

Bodily expression classification could also be referred to the previous section. However, **bodily expressions are largely encoded in movements** [74]. In [23], bodily expression classification is achieved through similar features like joints coordinates and velocities. Since the data they use are collected in lab settings, joints coordinates could be extracted easily with basic image processing.

To sum up, bodily expression recognition is still understudied compared with facial expression recognition. The main reason is the lack of efficient emotion-sensitive feature extraction approaches. Although deep learning has shown marvelous power in computer vision [66], it is not feasible to directly adopting the approach. Firstly, it usually takes a lot of data to train deep neural network (DNN) to avoid overfitting. From the perspective of learning theory [75],

DNN has huge VC dimension due to many parameters in the model. In this case, the size of samples should be large enough to ensure the model's generalizability. However, datasets for bodily expression recognition have very small size. Secondly, it is expensive to get accurate annotations from data. For popular topics like pose estimation, there are several large benchmark datasets. Another weakness of current bodily expression recognition research would be that most datasets have acted emotions in lab settings [23]. Natural/spontaneous emotions may tell a different story yet more applicable to real-world applications.

3.1.1.3 Speech Expression

There are two types of information encoded in speech, what is said and how it is said. For the first part, it is also referred as sentiment analysis in natural language processing (NLP). For the second part, classification is conducted on original audio signals.

Compared with images, audio signals are one dimension signal over time. According to [76], energy, pitch, spectrum and prosody in audio are highly related to emotion recognition. Audio signals are often segmented into several small intervals called frames. Local features could be extracted from those frames, whereas global features are extracted over a large interval of signals. Speech features could as be categorized as continuous speech features, voice quality features, spectral-based speech features, and nonlinear TEO-based features [76]. For speech expression recognition, HMM is often used with feature codebook constructed from all possible features [77]. A recent study has also shown good results with end-to-end deep neural network training [78].

In sentiment analysis, since words already have specific meanings, a naive thought would be assigning each word with certain emotion weights. Bag of words (BoW), a commonly used feature, counts frequencies of words' occurrence. More sophisticated features or models also take words' order and phrase into account. Graphical models are often used for classification [79].

3.1.1.4 Physiology Indicators

Picard *et al.* firstly investigate physiology indicators of emotions [80]. In this work, several sensors are set up to measure facial muscle tension, blood volume pressure, skin conductance and respiration. Data collected from sensors are directly used as features for predicting emotions after signal processing. In [81], electroencephalogram (EEG) data are used for emotion recognition. The spectral power of EEG signals in different bands is selected as features. While physiology indicators are less susceptible to social editing [1], it is intrusive and inconvenient to apply to real-world applications.

3.1.2 Application

Emotion recognition studies could have various real-world applications in areas like human-computer interaction, medicine, security, services, and game and entertainment. In [82], computer scientists and clinicians found out an automatic system with the similar framework with automatic affect recognition system could help predict if one has depression. In [83], Lee *et al.* explored speech emotion recognition in natural dialogue. Customer service and call center could adopt this kind of system for better services. Ji *et al.* introduce a framework to monitor human fatigue based on visual cues and contextual information [84]. Intelligent automobile system could use this system to avoid accidents. For example, drivers could get alerted if the system detects fatigue. Other possible applications could be concentration detection in education, lie detection in homeland security and so on. Although a certain kind of emotion is focused on a specific scenario usually, these systems often share similar framework. Therefore, research on general emotion recognition is very promising and easy to be transferred to a specific domain.

3.1.3 Comparison

Based on the previous introduction on each modality, advantages and difficulties for emotion recognition are listed in Table 3.1. From the table, it is easy to see that facial expression and speech expression have already had some efficient features extracted from the data. In contrast, it is still a challenging topic for bodily expression. However, recent deep learning techniques have brought the possibility of solving this task. Also, various studies have shown each modality has specific advantages for recognizing some certain emotions or dimension (valence, arousal, dominance).

3.2 Multimodal Emotion Recognition

As aforementioned, there are a lot of situations where one modality could not recognize emotions accurately. It is essential for systems to leverage as much information as possible to generate accurate outcomes. Before diving into approaches for combining multiple modalities, there are two issues that are worth discussing with. Firstly, although human perceive emotions from multiple modalities, the information have much redundancy most of the time. And for a real-world application, there is possibility of compromised/broken sensors. Hence, it is essential to make sure the systems have the similar redundancy without making multimodality a burden instead and different modalities should not be treated as mutually independent.

Modality	Advantages	Challenges
Face	established techniques for feature extraction; one of the most salient channel in human perception	susceptible to social editing; reliance on high resolution; not realistic to track at all time
Body	feasible to estimate from a distance; less susceptible to social editing; some affective states are easily conveyed through body movements; large amount of available data/video that contain bodily expression; suitable for data-driven methods	hard to extract features from visual cues; challenging task (pose estimation) in computer vision; limited bodily emotion datasets; computational expensive (even for testing) for DPM
Speech	established techniques for feature extraction in audio signal processing; directly provide contextual information	difficult to understand context (challenging task in NLP)
Physiology	especially promising with new kind of wearable device to collect health data; almost not susceptible to social editing	few studies on this modality; no available data currently

Table 3.1: Comparison on different emotion recognition modalities.

Secondly, emotions on different modalities may not be synchronous. As suggested in [85], facial expression and bodily expression are not synchronized and the apex of the emotions are not at the same timestamp. The similar issue could also occur among other modalities. Therefore, simply blending all the available information at the current frame may not be a smart choice.

Historically, studies on multimodal emotion recognition mainly focused on two levels of fusion, feature level, and decision level. In [23], feature level fusion is achieved by concatenating features vectors directly. Feature selection is used to reduce the feature dimension. Decision level fusion is achieved by voting. The classification result from each modality is combined by weights. Even though the fusion methods here are straightforward, the bi-modal recognition rate is higher than that of either modality alone. In [85], Gunes *et al.* proposed a rather sophisticated system for combining two modalities. The system takes features from face and body respectively into two temporal segment detection classifier for classifying apex frame. Final classification of emotion is subsequently conducted based on features from selected frame with fusion strategy as either feature level or decision level. The underlying idea in this work is that synchronization could be achieved at the feature level.

Investigation of data fusion could be traced back to Hall's work [86]. In this work, data fusion is classified into four levels. Data fusion strategy is usually needed to be designed for specific problem. Theoretically, data fusion is more of a system work than a machine learning work and information could be fused at any time during the classification process. In [87], Sebe *et al.* proposed to fuse multimodal information using probabilistic graphical models (PGM). PGMs like HMM and Bayesian networks could handle noisy features, temporal information, and missing values. One other good thing about PGM is that we can add a hidden state in the model if there are more inference relationship found. In [88], Greg *et al.* developed a supervised learning framework for Bayesian network based on AdaBoost. The framework is used for speech detection under unconstrained environments based on multimodal information from camera and microphone.

Apart from PGM, many other methodologies for data fusion have been reviewed recently in [89]. In [89], data fusion methodologies are classified as stage-based, feature level-based, and semantic meaning-based. Since emotion recognition task does not involve too many stages, stage-based data fusion may not be applicable. For feature level-based fusion, DNN could be used to extract a sparse representation of the raw data (also called autoencoder in deep learning). An example of this is video and audio autoencoder [90]. For semantic meaning-based fusion, ensemble and boosting methods are mentioned in [89] for multi-kernel learning. From the perspective of ensemble models, each modality could be viewed as a weak classifier. Therefore, it is possible to solve emotion recognition task in this fashion. Also, PGM is also referred as one of semantic meaning-based data fusion methodology in [89].

3.3 Crowdsourced Affect Annotation

Crowdsourcing from the Internet as a data collection process has been originally proposed to collect objective, non-affective data and received popularity in the machine learning community to acquire large-scale ground truth datasets. A school of data quality control methods has been proposed for crowdsourcing. Yet, crowdsourcing affect annotations is highly challenging due to the intertwined subjectivity of affect and uninformative participants. Very few studies report on the limitations and complexity of crowdsourcing affect annotations. As suggested by [20], inconsistency of crowdsourced affective data exists due to two factors. The first is the possible untrustworthiness of recruited participants due to the discrepancy between the purpose of study (collecting high quality data) and the incentive for participants (earning cash rewards). The second is the natural variability of humans' perceiving others' affective expressions, as was discussed earlier. [91] crowdsourced personality attributes. Although they analyzed agreements

among different participants, they did not conduct quality control, catering to the two stated factors in the crowdsourcing. [92], however, used an *ad hoc* gold standard to control annotation quality and each sample in the training set was only annotated once. [19] crowdsourced evoked emotions of stimuli images. Building on [19], [20] proposed a probabilistic model, named the GLBA, to jointly model each worker’s reliability and regularity — the two factors contributing to the inconsistent annotations — in order to improve the quality of affective data collected. Because the GLBA methodology is applicable for virtually any crowdsourced affective data, we use it for our data quality control pipeline as well.

3.4 Human Pose Representation

Human pose estimation has attracted increasing attention recently due to its usefulness in surveillance, augmented reality, self-driven car and so on. More importantly, it is an essential part in an automated bodily expression recognition pipeline. As shown in Figure 3.3, in this section, we review previous work on three different human pose representation: 2D pose, 3D pose and human pose mesh.

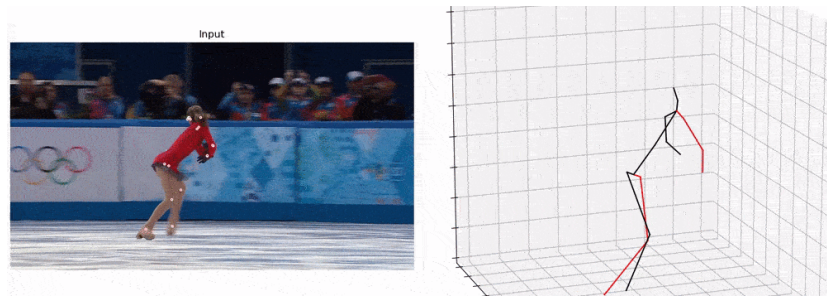
3.4.1 2D Pose Estimation

To define 2D pose estimation task well, it is essential to have a representation of body for further estimation. There are several kinds of models for modeling a human body in pose estimation. The main differences between these models lie in the choice of joints, the way to represent the body, and degree of freedom in them. As shown in Figure 3.4, cardboard people and pictorial structures treat limbs as several rigid polygonal parts and those parts are articulated through joints. However, contour people and deformable structures models capture not only configurations of body but also shape characteristics of body parts. In [3], Zuffi *et al.* argues that body representation is critically important. However, recent pose estimation researches are mostly based on models similar to pictorial structures.

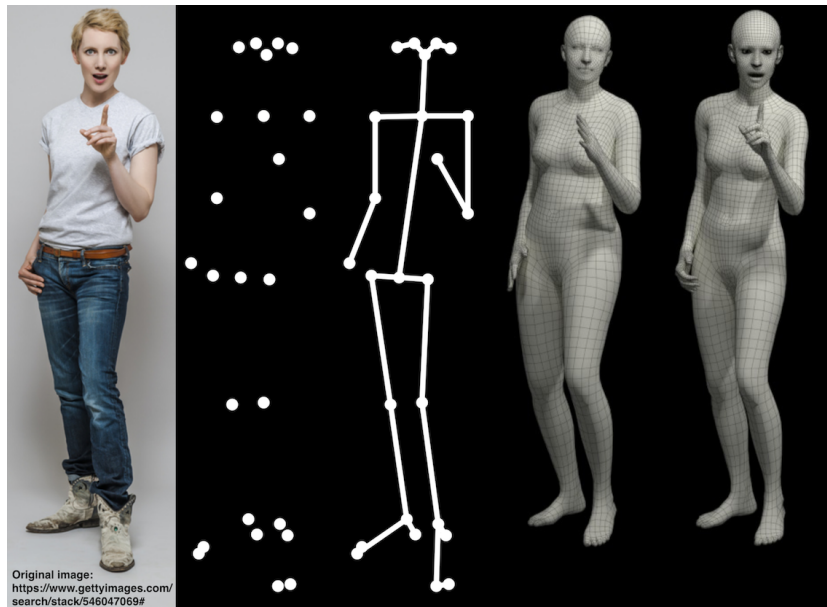
2D Pose estimation from still images has been explored a lot since Felzenszwalb *et al.* proposed deformable part model (DPM) for object detection [69]. For DPM-based approaches, the best performance of pose estimation from still images is kept by Yang *et al.* [70]. In DPM models, rigid parts are articulated through joints with spring-like constraints. The model extracts dense HOG features from image pyramid and minimizes the energy that reflects whether the part could match with the root part. The results in [70] indicate that the accuracies on head, shoulders, and torso detection are rather high. However, the algorithm fails constantly



(1) 2D pose. Image from [27].



(2) 3D pose. Image from [93]



(3) Human pose mesh. Image from [94]

Figure 3.3: Three different human pose representation: 2D pose, 3D pose and human pose mesh.

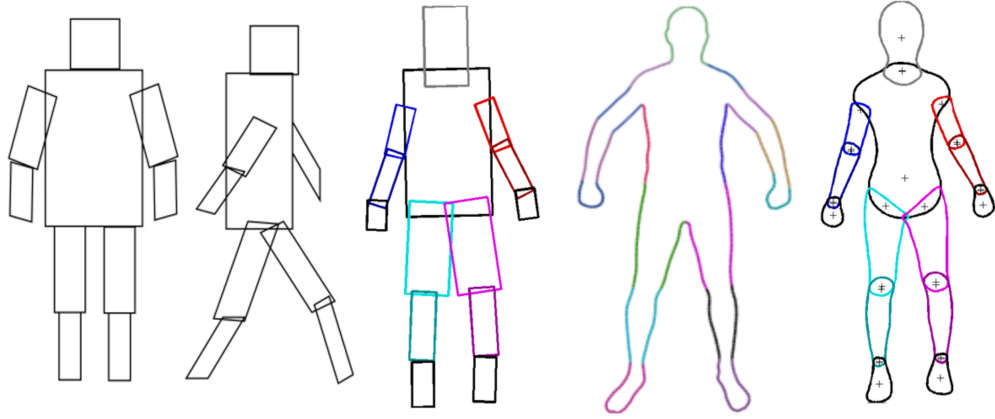


Figure 3.4: 2D models of human shape. From the left: Cardboard People (in two viewpoints), Pictorial Structures (PS), Contour People (CP) and Deformable Structures (DS). Image from Figure 2 of [3].

in detecting lower arms.

Recent deep learning based methods [27] formulate pose estimation as a detection problem, *i.e.*, detecting a predefined collection of landmarks of human body. The reasons for the recent significant progress in 2D pose estimation are two-fold. First, heatmap representation of joints are easier to regress for a convolutional neural network. Second, an encoder-decoder architecture enables more efficient learning process [95]. One significant drawback of this landmark-based 2D pose representation is that both depth and appearance information is ignored.

3.4.2 3D Pose Estimation

Instead of only estimating 2D coordinates of human landmarks, 3D pose estimation lifts the estimation of 2D coordinates into 3D. There are two common pipelines for solving the problem in the literature. The first one uses 2D pose estimation result as input and regress 3D pose directly [28]. The second one jointly model 2D pose estimation and 3D pose estimation [96]. This pose representation cannot keep the appearance information and subtle expressive pose cannot be faithfully recovered consequently. One significant obstacle in this line of research is lack of 3D pose data.

3.4.3 Human Pose and Shape Reconstruction

Human pose can also be represented with human pose mesh, *i.e.*, point clouds of the surface of a person. One representative model is the SMPL model [4]. SMPL model is a parametric graphics model that takes two group of parameters as input. The first group of parameters controls how each two parts of the body are articulated through a joint and the second group of parameters controls the body shape of a person, *i.e.*, tall or short. One of the challenges in this research is the limitation of ground truth, *i.e.*, parameters of 3D human model (SMPL [4]). In fact, the parameters of these graphics model cannot be directly measured or observed. Therefore, the optimization goal or the supervision signal from in-the-wild data usually comes from other observable characteristics, *e.g.*, 2D keypoints, silhouettes and semantic parts [97–100]. On the other hand, the advantage of this representation is that it could capture detailed information of a subtle pose.

3.5 Summary

In this chapter, automated emotion recognition pipeline for each unimodal is reviewed. Interestingly, we found there is no attempt to leverage movement notation system in bodily expression recognition research. We think there are two reasons for this. First, clear correspondence between specific movements and emotions is not found yet. Second, automated movement notation extraction is a challenging task in computer vision research although pose estimation results have been improved significantly these years.

Chapter 4 | The BoLD Dataset

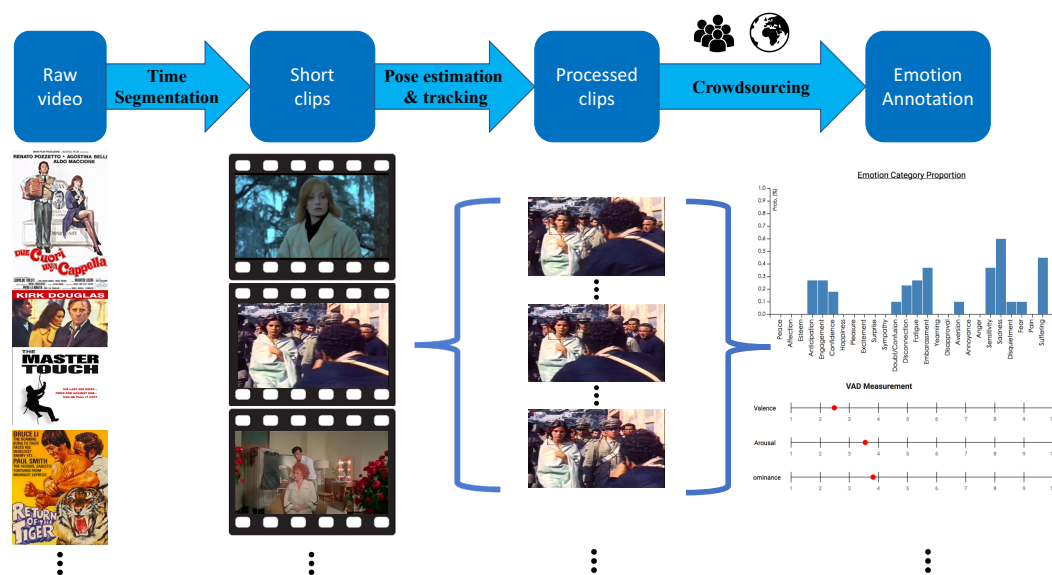


Figure 4.1: Overview of our data collection pipeline. The process involves crawling movies, segmenting them into clips, estimating the poses, and emotion annotation.

4.1 Dataset Construction

The dataset construction process, detailed below, consists of three stages: movie selection and time segmentation, pose estimation and tracking, and emotion annotation. Fig. 4.1 illustrates our dataset construction pipeline. We chose the movies included in a public dataset, the AVA

The work has been published as “ARBEE: Towards Automated Recognition of Bodily Expression of Emotion In the Wild”, *International Journal of Computer Vision*, vol. 128, no. 1, pp. 1-25, 2020 [101].

dataset [102], which contains a list of YouTube movie IDs. To respect the copyright of the movies, we provide the movie ID in the same way as in the AVA dataset when the data is shared to the research community. Any raw movies will be kept only for feature extraction and research in the project and will not be distributed. Given raw movies crawled from Youtube, we first partitioned each into several short scenes before using other vision-based methods to locate and track each person across different frames in the scene. To facilitate tracking, the same person in each clip was marked with a unique ID number. Finally, we obtained emotion annotations of each person in these ID-marked clips by employing independent contractors (to be called participants hereafter) from the online crowdsourcing platform, the Amazon Mechanical Turk (AMT).

4.1.1 Movie Selection and Time Segmentation

The Internet has vast *natural* human-to-human interaction videos, which serves as a rich source for our data. A large collection of video clips from daily lives is an ideal dataset for developing affective recognition capabilities because they match closely with our common real-world situations. However, a majority of those user-uploaded, in-the-wild videos suffer from poor camera perspectives and may not cover a variety of emotions. We consider it beneficial to use movies and TV shows, *e.g.*, reality shows or uploaded videos in social media, that are unconstrained but offer highly interactive and emotional content. Movies and TV shows are typically of high quality in terms of filming techniques and the richness of plots. Such shows are thus more representative in reflecting characters’ emotional states than some other categories of videos such as DIY instructional videos and news event videos, some of which were collected recently [103, 104]. In this work, we have crawled 150 movies (220 hours in total) from YouTube by the video IDs curated in the AVA dataset [102]. Figure 4.2 lists all movie titles.

Movies are typically filmed so that shots in one scene demonstrate characters’ specific activities, verbal communication, and/or emotions. To make these videos manageable for further human annotation, we partition each video into short video clips using the kernel temporal segmentation (KTS) method [105]. KTS detects shot boundary by keeping variance of visual descriptors within a temporal segment small. Shot boundary can be either a change of scene or a change of camera perspective within the same scene. To avoid confusion, we will use the term scene to indicate both cases.

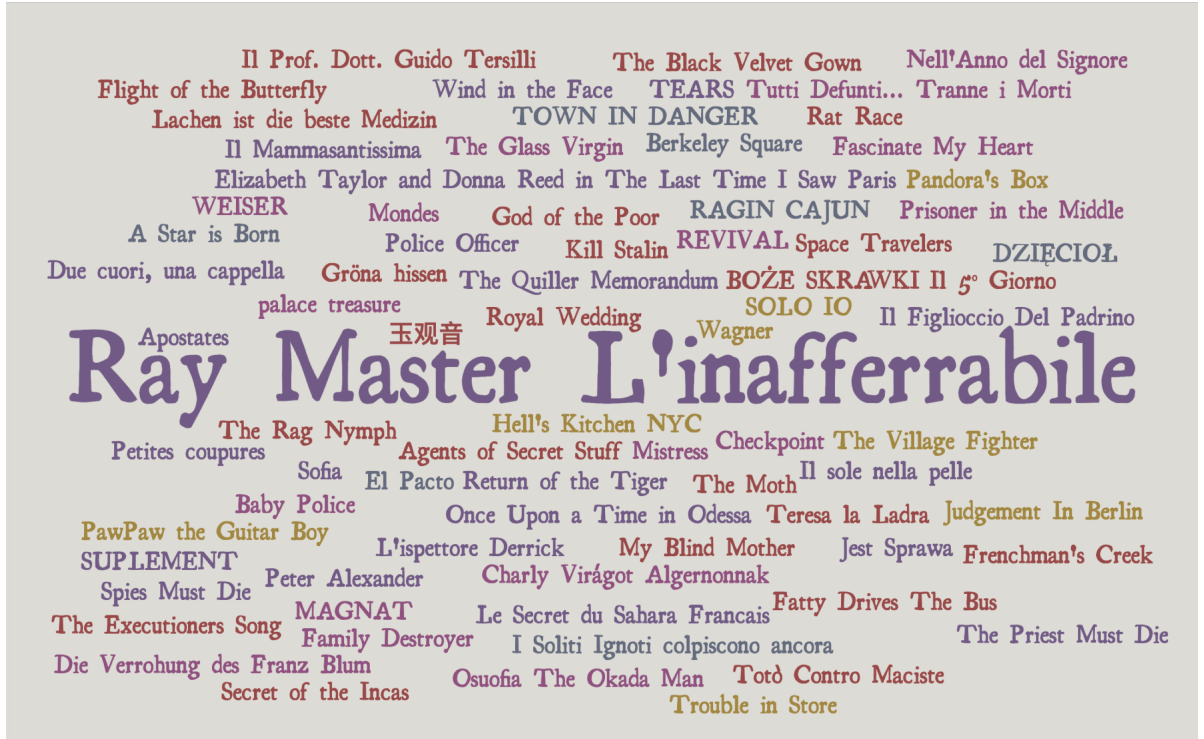


Figure 4.2: Selected movie titles.

4.1.2 Pose Estimation and Tracking

We adopted an approach to detect human body landmarks and track each character at the same time (Fig. 4.3). Because not all short clips contain human characters, we removed those clips without humans via pose estimation [27]. Each clip was processed by a pose estimator¹ frame-by-frame to acquire human body landmarks. Different characters in one clip correspond to different samples. Each character in the clip is marked as a different sample. To make the correspondence clear, we track each character and designate them with a unique ID number. Specifically, tracking was conducted on the upper-body bounding box with the Kalman Filter and Hungarian algorithm as the key component [106]². In our implementation, the upper-body bounding box was acquired with the landmarks on face and shoulders. Empirically, to ensure reliable tracking results when presenting to the annotators, we removed short trajectories that had less than 80% of the total frames.

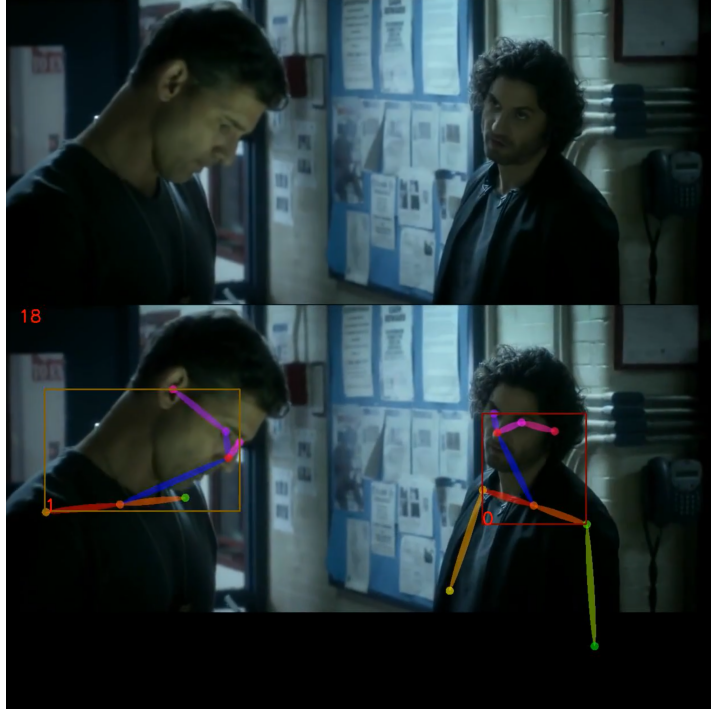


Figure 4.3: A frame in a video clip, with different characters numbered with an ID (*e.g.*, 0 and 1 at the bottom left corner of red bounding boxes) and the body and/or facial landmarks detected (indicated with the stick figure).

4.1.3 Emotion Annotation

Following the above steps, we generated 122,129 short clips from these movies. We removed facial close-up clips using results from pose estimation. Concretely, we included a clip in our annotation list if the character in it has at least three visible landmarks out of the six upper-body landmarks, *i.e.*, wrists, elbows, and shoulders on both body sides (left and right). We further select those clips with between 100 and 300 frames for manual annotation by the participants. An identified character with landmark tracking in a single clip is called an *instance*. We have curated a total of 48,037 instances for annotation from a total of 26,164 video clips.

We used the AMT for crowdsourcing emotion annotations of the 48,037 instances. For each Human Intelligence Task (HIT), a human participant completes emotion annotation assignments for 20 different instances. Each of which was drawn randomly from the instance pool. Each instance is expected to be annotated by five different participants.

We asked human annotators to finish four annotation tasks per instance. Fig. 4.4 shows screenshots of our crowdsourcing website design. As a first step, participants must check if

¹https://github.com/CMU-Perceptual-Computing-Lab/caffe_rtpose

²<https://github.com/abewley/sort>



(1) video data quality check



(2) categorical emotion labeling



(3) dimensional emotion and demographic labeling



(4) frame range identification

Figure 4.4: The web-based crowdsourcing data collection process. Screenshots of the four steps are shown. For each video clip, participants are directed to go through a sequence of screens with questions step-by-step.

the instance is corrupted. An instance is considered corrupted if landmark tracking of the character is not consistent or the scene is not realistic in daily life, such as science fiction scenes. If an instance is not corrupted, participants are asked to annotate the character's emotional expressions according to both categorical emotions and dimensional emotions (*i.e.*, valence, arousal, dominance (VAD) in dimensional emotion state model mehrabian1980basic). For categorical emotions, we used the list in [92], which contains 26 categories and is a superset of the six basic emotions [33]. Participants are asked to annotate these categories in the way of multi-label binary classifications. For each dimensional emotion, we used integers that scales from 1 to 10. These annotation tasks are meant to reflect the truth revealed in the visual and audio data — movie characters' emotional expressions — and do not involve the participants' emotional feelings. In addition to these tasks, participants are asked to specify a time interval (*i.e.*, the start and end frames) over the clip that best represents the selected emotion(s) or has led to their annotation. Characters' and participants' demographic information (gender, age,

and ethnicity) is also annotated/collected for complementary analysis. Gender categories are male and female. Age categories are defined as kid (aged up to 12 years), teenager (aged 13-20), and adult (aged over 20). Ethnicity categories are American Indian or Alaska Native, Asian, African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, White, and Other.

The participants are permitted to hear the audio of the clip, which can include a conversation in English or some other language. While the goal of this research is to study the computability of body language, we allowed the participants to use all sources of information (facial expression, body movements, sound, and limited context) in their annotation in order to obtain as high accuracy as possible in the data collected. Additionally, the participants can play the clip back-and-forth during the entire annotation process for that clip.

To sum up, we crowdsourced the annotation of categorical and dimensional emotions, time interval of interest, and character demographic information.

4.1.4 Annotation Quality Control

Quality control has always been a necessary component for crowdsourcing to identify dishonest participants, but it is much more difficult for affect data. Different people may not perceive affect in the same way, and their understanding may be influenced by their cultural background, current mood, gender, and personal experiences. An honest participant could also be uninformative in affect annotation, and consequently, their annotations can be poor in quality. In our study, the variance in acquiring affects usually comes from two kinds of participants, *i.e.*, dishonest ones, who give useless annotations for economic motivation, and exotic ones, who give inconsistent annotations compared with others. Note that exotic participants come with the nature of emotion, and annotations from exotic participants could still be useful when aggregating final ground truth or investigating cultural or gender effects of affect. In our crowdsourcing task, we want to reduce the variance caused by dishonest participants. In the meantime, we do not expect too many exotic participants because that would lead to low consensus.

Using gold standard examples is a common practice in crowdsourcing to identify uninformative participants. This approach involves curating a set of instances with known ground truth and removing those participants who answer incorrectly. For our task, however, this approach is not as feasible as in conventional crowdsourcing tasks such as image object classification. To accommodate subjectivity of affect, gold standard has to be relaxed to a large extent. Consequently, the recall of dishonest participants is lower.

To alleviate the aforementioned dilemma, we used four complementary mechanisms for quality control, including three online approaches (*i.e.*, analyzing while collecting the data) and

an offline one (*i.e.*, post-collection analysis). The online approaches are participant screening, annotation sanity check, and relaxed gold standard test, while the offline one is reliability analysis.

- *Participant screening.* First-time participants in our HIT must take a short empathy quotient (EQ) test [107]. Only those who have above-average EQ are qualified. This approach aims to reduce the number of exotic participants from the beginning.
- *Annotation sanity check.* During the annotation process, the system checks consistency between categorical emotion and dimensional emotion annotations as they are entered. Specifically, we expect an “affection”, “esteem”, “happiness”, or “pleasure” instance to have an above-midpoint valence score; a “disapproval”, “aversion”, “annoyance”, “anger”, “sensitivity”, “sadness”, “disquietment”, “fear”, “pain”, or “suffering” instance to have a below-midpoint valence score; a “peace” instance to have a below-midpoint arousal score; and an “excitement” instance to have an above-midpoint arousal score. As an example, if a participant chooses “happiness” and a valence rating between 1 and 5 (out of 10) for an instance, we treat the annotation as inconsistent. In each HIT, a participant fails this annotation sanity check if there are two inconsistencies among twenty instances.
- *Relaxed gold standard test.* One control instance (relaxed gold standard) is randomly inserted in each HIT to monitor the participant’s performance. We collect control instances in our trial run within a small trusted group and choose instances with very high consensus. We manually relax the acceptable range of each control instance to avoid false alarm. For example, for an indisputable sad emotion instance, we accept an annotation if valence is not higher than 6. An annotation that goes beyond the acceptable range is treated as failing the gold standard test. We selected nine control clips and their relaxed annotations as the gold standard. We did not use more control clips because the average number of completed HITs per participant is much less than nine and the gold standard is rather relaxed and inefficient in terms of recall.
- *Reliability analysis.* To further reduce the noise introduced by dishonest participants, we conduct reliability analysis over all participants. We adopted the method by [20] to properly handle the intrinsic subjectivity in affective data. Reliability and regularity of participants are jointly modeled. Low-reliability-score participant corresponds to dishonest participant, and low-regularity participant corresponds to exotic participant. This method was originally developed for improving the quality of dimensional annotations based on modeling the agreement multi-graph built from all participants and their

annotated instances. For each dimension of VAD, this method estimates participant i 's reliability score, *i.e.*, r_i^v, r_i^a, r_i^d . According to [20], the valence and arousal dimensions are empirically meaningful for ranking participants' reliability scores. Therefore, we ensemble the reliability score as $r_i = (2r_i^v + r_i^a)/3$. We mark participant i as failing in reliability analysis if r_i is less than $\frac{1}{3}$ with enough effective sample size.

Based on these mechanisms, we restrain those participants deemed 'dishonest.' After each HIT, participants with low performance are blocked for one hour. Low-performance participant is defined as either failing the annotation sanity check or the relaxed gold standard test. We reject the work if it shows low performance and fails in the reliability analysis. In addition to these constraints, we also permanently exclude participants with a low reliability score from participating our HITs again.

4.1.5 Annotation Aggregation

Whenever a single set of annotations is needed for a clip, proper aggregation is necessary to obtain a consensus annotation from multiple participants. The Dawid-Skene method [108], which is typically used to combine noisy categorical observations, computes an estimated score (scaled between 0 and 1) for each instance. We used the method to aggregate annotations on each categorical emotion annotation and categorical demographic annotation. Particularly, we used the notation s_i^c to represent the estimated score of the binary categorical variable c for the instance i . We set a threshold of 0.5 for these scores when binary categorical annotation is needed. For dimensional emotion, we averaged the set of annotations for a clip with their annotators' reliability score [20]. Considering a particular instance, suppose it has received n annotations. The score s_i^d is annotated by participant i with reliability score r_i for dimensional emotion d , where $i \in \{1, 2, \dots, n\}$ and $d \in \{V, A, D\}$ in the VAD model. The final annotation is then aggregated as

$$s^d = \frac{\sum_{i=1}^n r_i s_i^d}{10 \sum_{i=1}^n r_i} . \quad (4.1)$$

In the meantime, instance confidence according to the method by [20] is defined as

$$c = 1 - \prod_{i=1}^n (1 - r_i) . \quad (4.2)$$

Note that we divided the final VAD score by 10 so that the data ranges between 0 and 1. Our final dataset to be used for further analysis retained only those instances with confidence higher than 0.95.

Our website sets a default value for the start frame (0) and the end frame (total frame number of the clip) for each instance. Among the data collected, there were about a half annotations that have non-default values, which means a portion of the annotators either considered the whole clip as the basis for their annotations or did not finish the task. For each clip, we selected the time-interval entered by the participant with the highest reliability score as the final annotation for the clip.

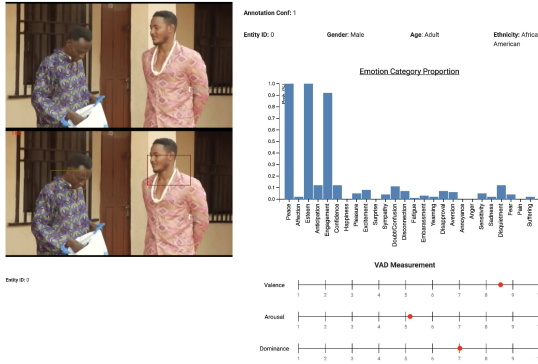
4.2 Dataset Statistics

We report relevant dataset statistics. We used state-of-the-art statistical techniques to validate our quality control mechanisms and thoroughly understand the consensus level of our verified data labels. Because human perceptions of a character’s emotions naturally varies across participants, we do not expect absolute consensus for collected labels. In fact, it is nontrivial to quantitatively understand and measure the quality of such affective data.

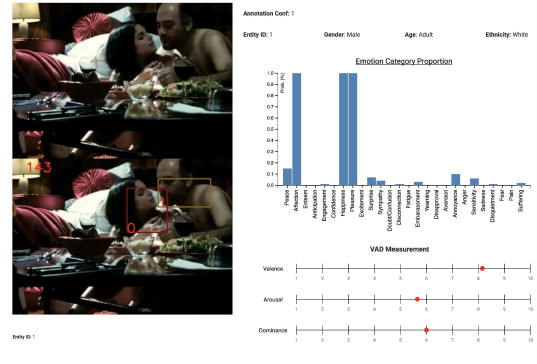
4.2.1 Annotation Distribution and Observations

We have collected annotations for 13,239 instances. The dataset continues to grow as more instances and annotations are added. Fig. 4.5 shows some high-confidence instances in our dataset. Figs. 4.6, 4.7, and 4.8 show the distributions of categorical emotion, dimensional emotion, and demographic information, respectively. For each categorical emotion, the distribution is highly unbalanced. For dimensional emotion, the distributions of three dimensions are Gaussian-like, while valence is right-skewed and dominance is left-skewed. Character demographics is also unbalanced: most characters in our movie-based dataset are male, white, and adult. We partition all instances into three sets: the training set ($\sim 70\%$, 9222), the validation set ($\sim 10\%$, 1153), and the testing set (20%, 2864). Our split protocol ensured that clips from the same raw movie video belong to the same set so that subsequent evaluations can be conducted faithfully.

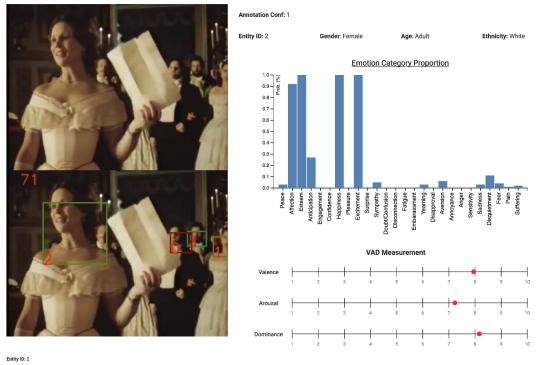
We observed interesting correlations between pairs of categorical emotions and pairs of dimensional emotions. Fig. 4.9 shows correlations between each pair of emotion categories. Categorical emotion pairs such as pleasure and happiness (0.57), happiness and excitement (0.40), sadness and suffering (0.39), annoyance and disapproval (0.37), sensitivity and sadness (0.37), and affection and happiness (0.35) show high correlations, matching our intuition. Correlations between dimensional emotions (valence and arousal) are weak (0.007). Because these two dimensions were designed to indicate independent characteristics of emotions, weak



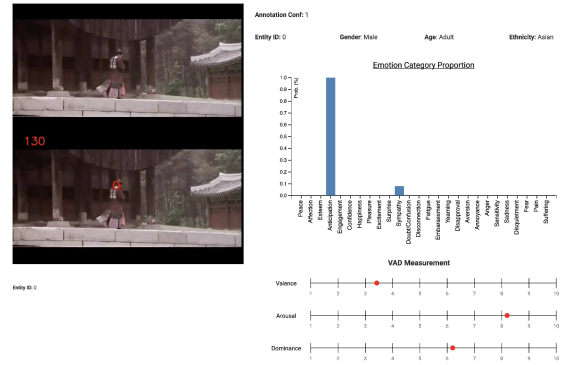
(1) peace



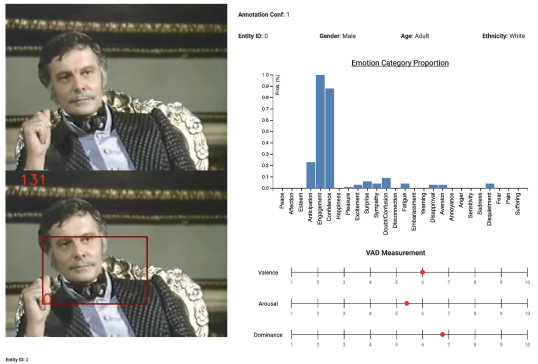
(2) affection



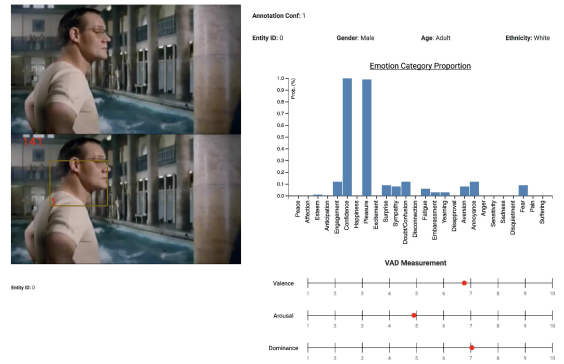
(3) esteem



(4) anticipation



(5) engagement

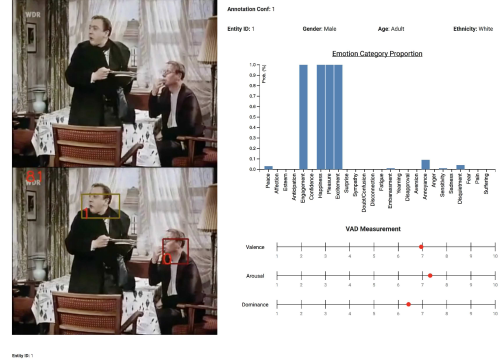


(6) confidence

Figure 4.5: Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion. To be continued on the next page.



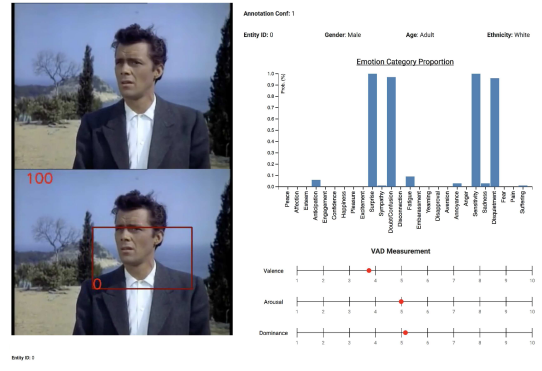
(7) happiness



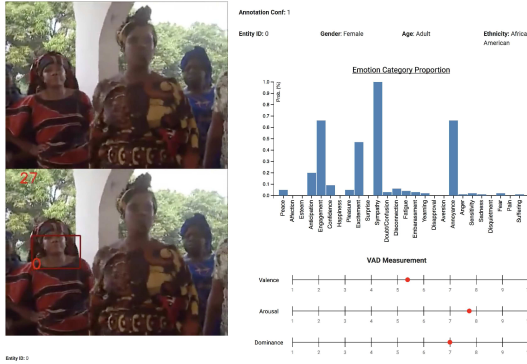
(8) pleasure



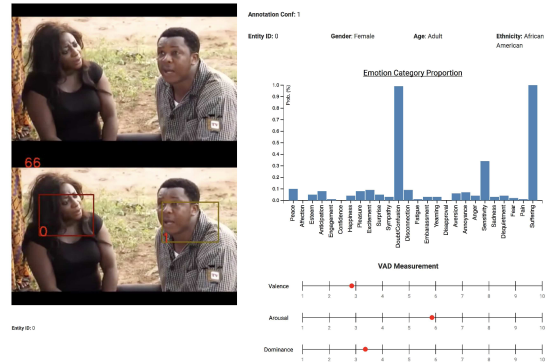
(9) excitement



(10) surprise

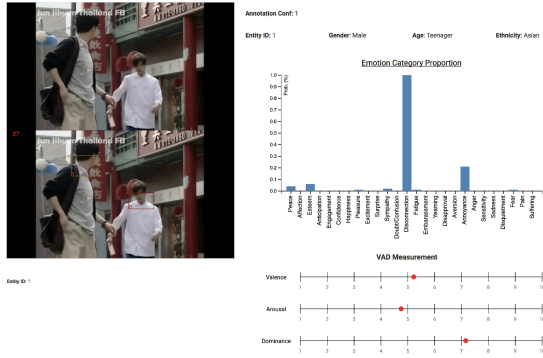


(11) sympathy

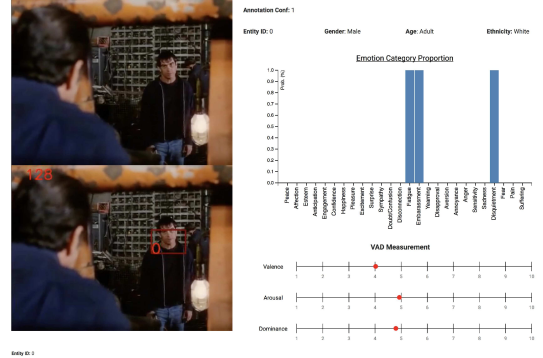


(12) doubt/confusion

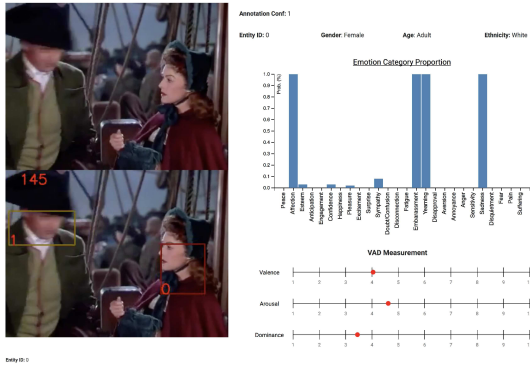
Figure 4.5: (Continued from the previous page.) Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances (27 and 28) that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion.



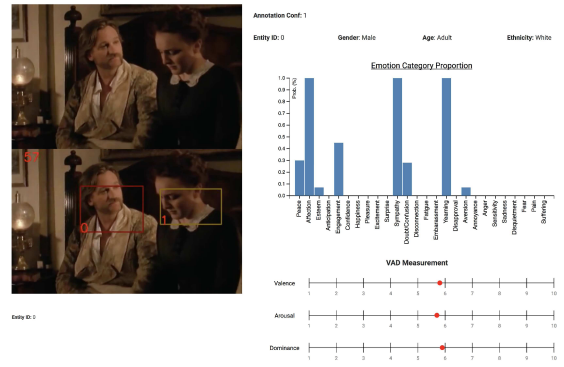
(13) disconnection



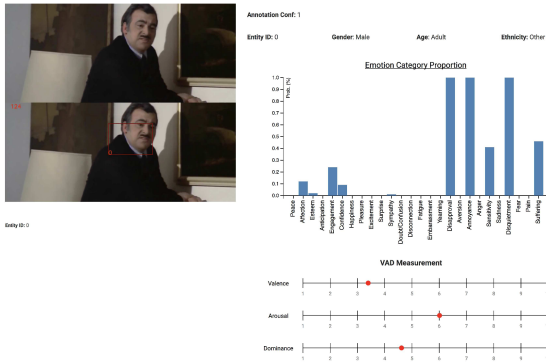
(14) fatigue



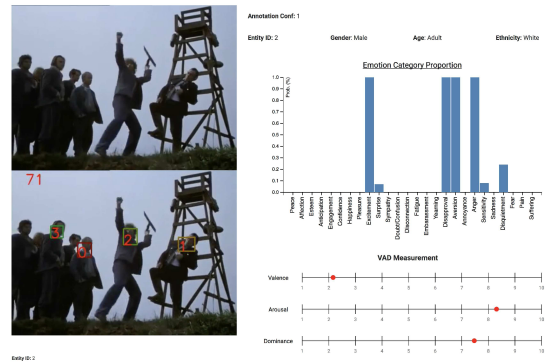
(15) embarrassment



(16) yearning



(17) disapproval

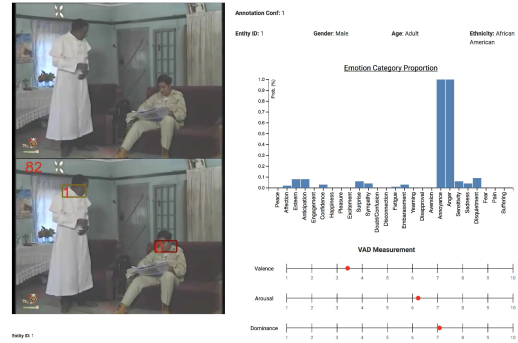


(18) aversion

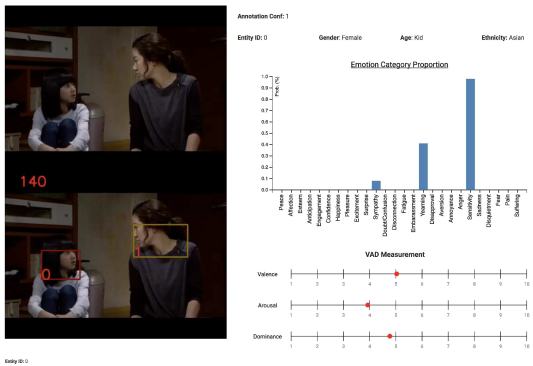
Figure 4.5: (Continued from the previous page.) Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances (27 and 28) that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion.



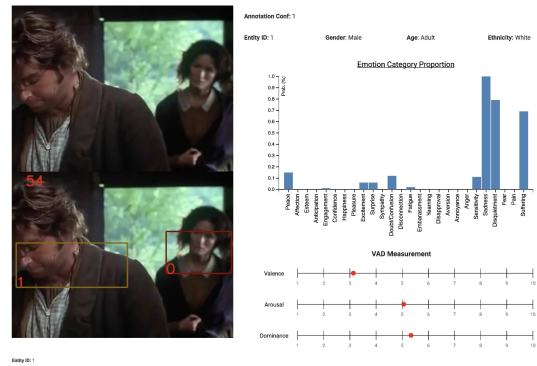
(19) annoyance



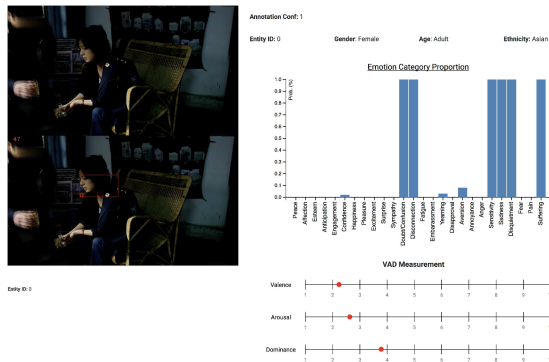
(20) anger



(21) sensitivity



(22) sadness

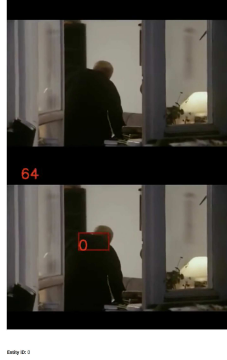


(23) disquietment

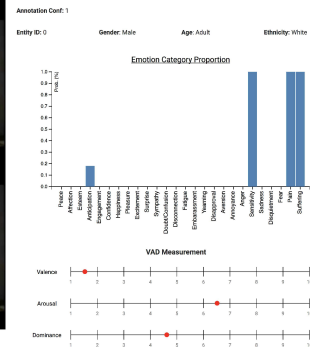


(24) fear

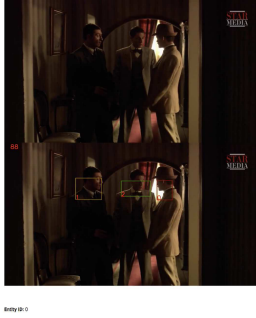
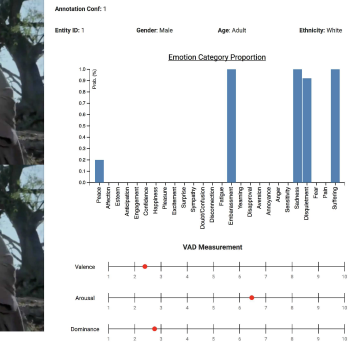
Figure 4.5: (Continued from the previous page.) Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances (27 and 28) that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion.



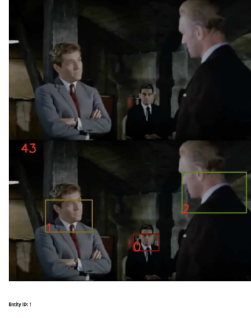
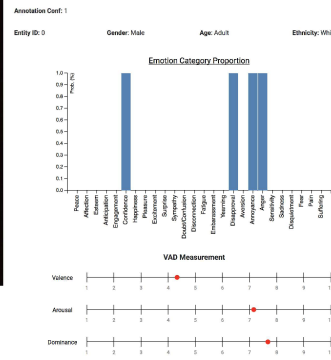
(25) pain



(26) suffering



(27) quality control



(28) quality control

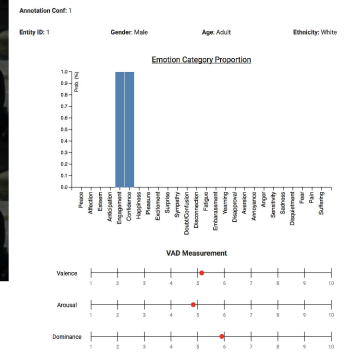


Figure 4.5: (Continued from the previous page.) Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances (27 and 28) that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c , demographics of the character, and aggregated categorical and dimensional emotion.

correlations among them confirm their validity. However, correlations between valence and dominance 0.359, and between arousal and dominance (0.356) are high. This finding is evidence that dominance is not a strictly independent dimension in the VAD model.

We also observed sound correlations between dimensional and categorical emotions. Valence shows strong positive correlations with happiness (0.61) and pleasure (0.51), and strong negative correlations with disapproval (-0.32), sadness (-0.32), annoyance (-0.31), and disquitement (-0.32). Arousal shows positive correlations with excitement (0.25) and anger (0.31), and negative correlations with peace (-0.20), and disconnection (-0.23). Dominance shows strong correlation with confidence (0.40), and strong negative correlation with doubt/-confusion (-0.23), sadness (-0.28), fear (-0.23), sensitivity (-0.22), disquitement (-0.24),

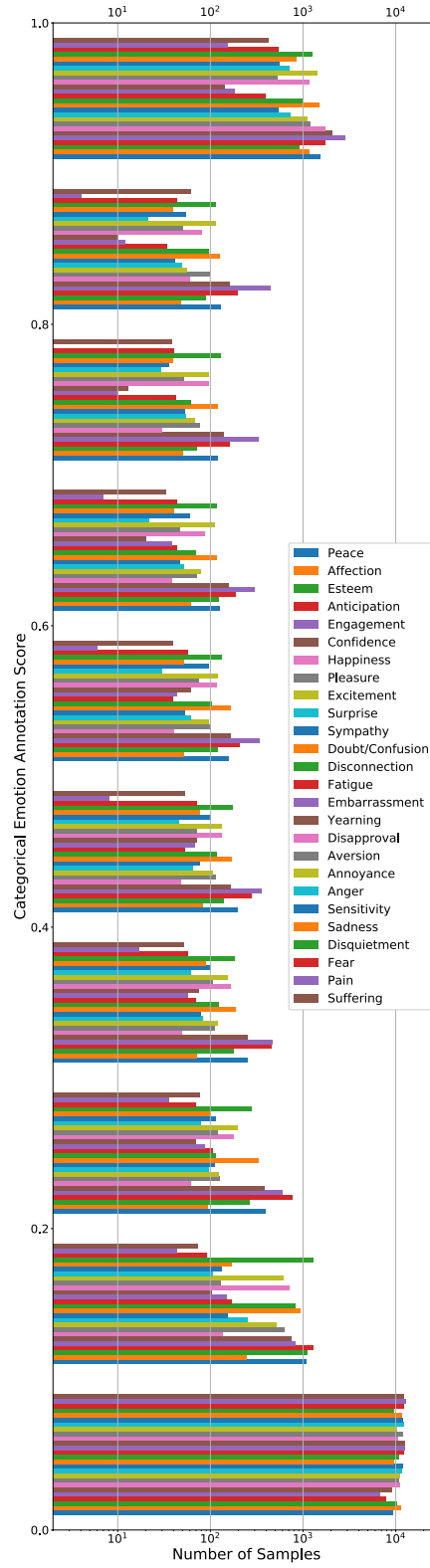


Figure 4.6: Distributions of the 26 different categorical emotions.

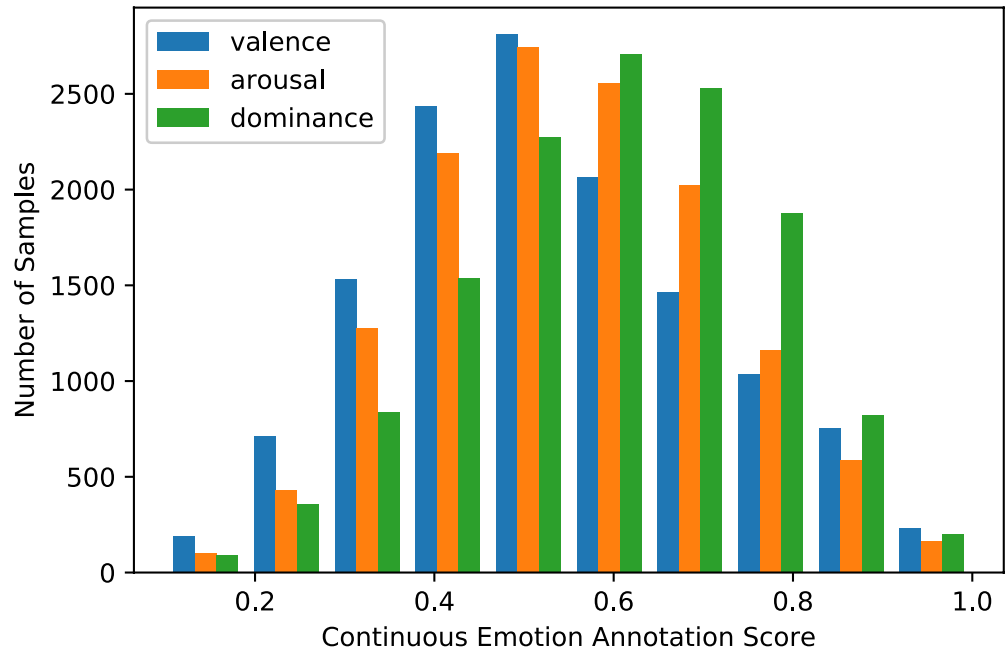


Figure 4.7: Distributions of the three dimensional emotion ratings: valence, arousal, and dominance.

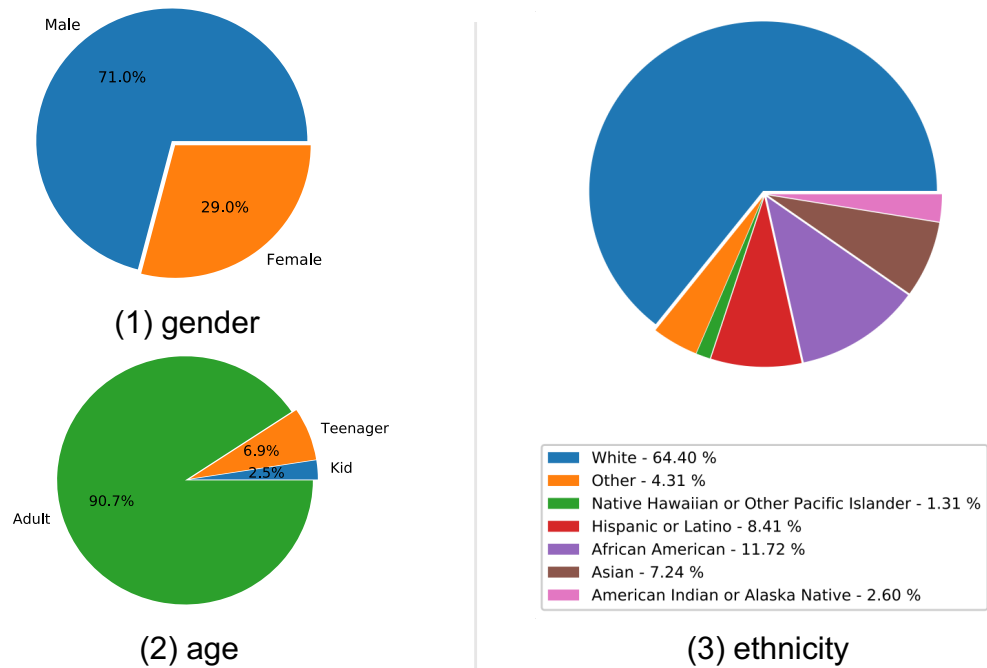


Figure 4.8: Demographics of characters in our dataset.

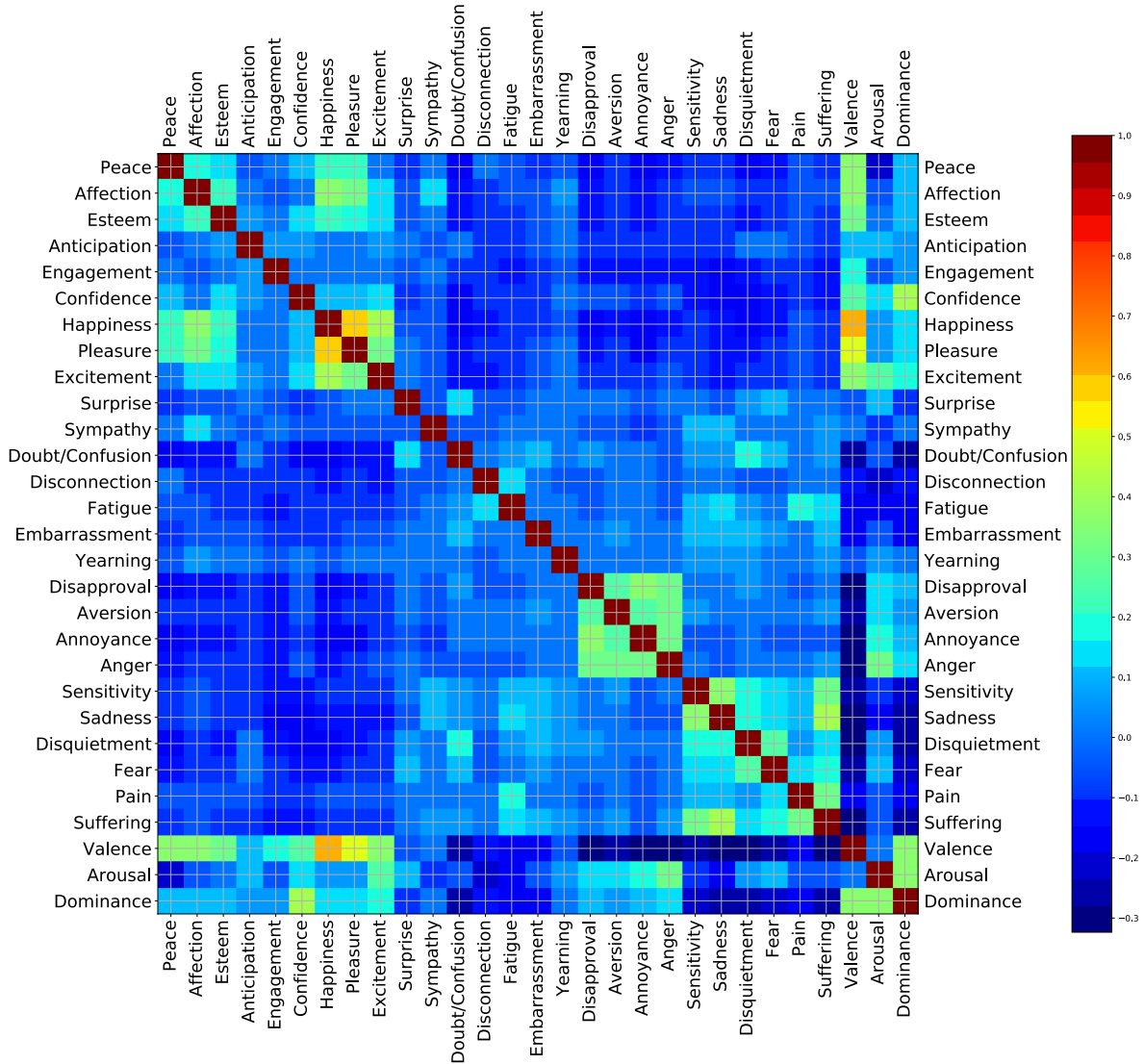


Figure 4.9: Correlations between pairs of categorical or dimensional emotions, calculated based on the BoLD dataset.

and suffering (-0.25). All of these correlations match with our intuition about these emotions.

4.2.2 Annotation Quality and Observations

We computed Fleiss’ Kappa score (κ) for each categorical emotion and categorical demographic information to understand the extent and reliability of agreement among participants. Perfect agreement leads to a score of one, while no agreement leads to a score less than or equal to zero. Table 4.1 shows Fleiss’ Kappa [109] among participants on each categorical emotion and categorical demographic information. κ is computed on all collected annotations for each

Table 4.1: Agreement among participants on categorical emotions and characters’ demographic information.

Category	κ	filtered κ	Category	κ	filtered κ
Peace	0.132	0.148	Affection	0.262	0.296
Esteem	0.077	0.094	Anticipation	0.071	0.078
Engagement	0.110	0.126	Confidence	0.166	0.183
Happiness	0.385	0.414	Pleasure	0.171	0.200
Excitement	0.178	0.208	Surprise	0.137	0.155
Sympathy	0.114	0.127	Doubt/Confusion	0.127	0.141
Disconnection	0.125	0.140	Fatigue	0.113	0.131
Embarrassment	0.066	0.085	Yearning	0.030	0.036
Disapproval	0.140	0.153	Aversion	0.075	0.087
Annoyance	0.176	0.197	Anger	0.287	0.307
Sensitivity	0.082	0.097	Sadness	0.233	0.267
Disquietment	0.110	0.125	Fear	0.193	0.214
Pain	0.273	0.312	Suffering	0.161	0.186
Average	0.154	0.173			
Gender	0.863	0.884	Age	0.462	0.500
Ethnicity	0.410	0.466			

category. For each category, we treated it as a two-category classification and constructed a subject-category table to compute Fleiss’ Kappa. By filtering out those with low reliability scores, we also computed filtered κ . Note that some instances may have less than five annotations after removing annotations from low-reliability participants. We edited the way to compute p_j , defined as the proportion of all assignments which were to the j -th category. Originally, it should be

$$p_j = \frac{1}{N} \sum_{i=1}^N \frac{n_{ij}}{n}, \quad (4.3)$$

where N is the number of instances, n_{ij} is the number of ratings annotators have assigned to the j -th category on the i -th instance, and n is the number of annotators per instance. In our filtered κ computation, n varies for different instances and we denote the number of annotators

for instance i as n_i . Then Eq. (4.3) is revised as:

$$p_j = \frac{1}{N} \sum_{i=1}^N \frac{n_{ij}}{n_i} . \quad (4.4)$$

Filtered κ is improved for each category, even for those objective category like gender, which also suggests the validity of our offline quality control mechanism. Note that our reliability score is computed over dimensional emotions, and thus the offline quality control approach is complementary. As shown in the table, affection, anger, sadness, fear, and pain have fair levels of agreement ($0.2 < \kappa < 0.4$). Happiness has moderate level of agreement ($0.4 < \kappa < 0.6$), which is comparable to objective tasks such as age and ethnicity. This result indicates that humans are mostly consistent in their sense of happiness. Other emotion categories fall into the level of slight agreement ($0 < \kappa < 0.2$). Our κ score of demographic annotation is close to previous studies reported in [91]. Because the annotation is calculated from the same participant population, κ also represents how difficult or subjective the task is. Evidently gender is the most consistent (hence the easiest) task among all categories. The data confirms that emotion recognition is both challenging and subjective even for human beings with sufficient level of EQ. Participants in our study passed an EQ test designed to measure one’s ability to sense others’ feelings as well as response to others’ feelings, and we suspect that individuals we excluded due to a failed EQ test would likely experience greater difficulty in recognizing emotions.

For dimensional emotions, we computed both across-annotation variances and within-annotation variances. The variances across all annotations are 5.87, 6.66, and 6.40 for valence, arousal, and dominance, respectively. Within-instance variances (over different annotators) is computed for each instance and the means of these variances are 3.79, 5.24, and 4.96, respectively. Notice that for the dimensions, the variances are reduced by 35%, 21%, and 23%, respectively, which illustrates human performance at reducing variance given concrete examples. Interestingly, participants are better at recognizing positive and negative emotions (*i.e.* valence) than in other dimensions.

4.2.3 Human Performance

We explored the difference between low-performance participants and low reliability-score participants. As shown in Fig. 4.10, low-performance participants shows lower reliability score by average. While a significantly large number of low-performance participants have rather high reliability scores, most non-low-performance participants have reliability scores larger

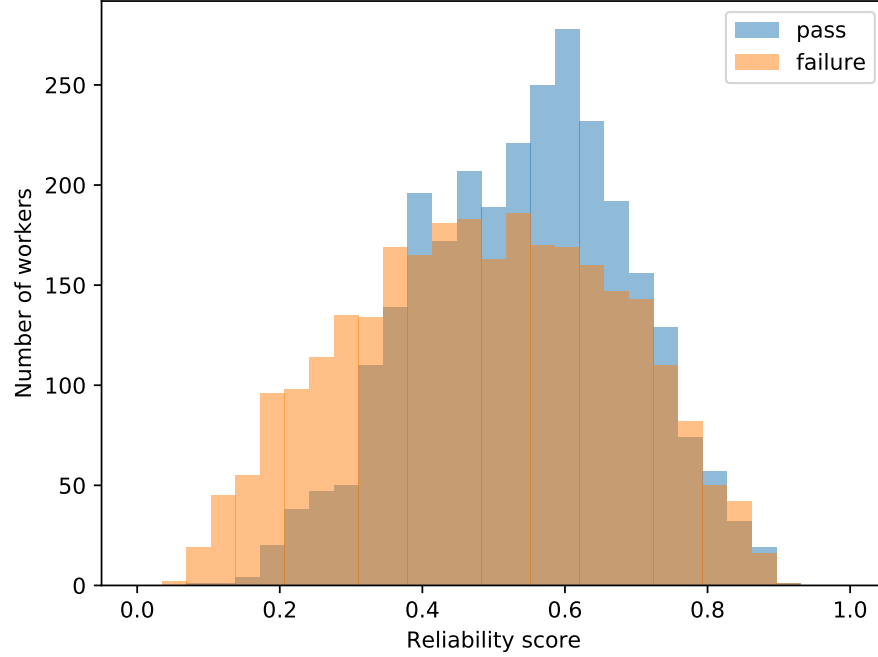
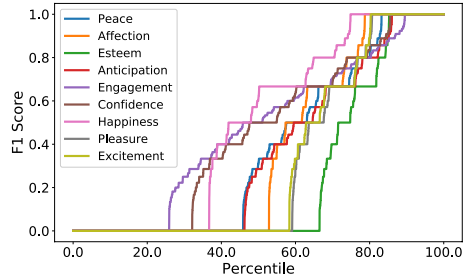


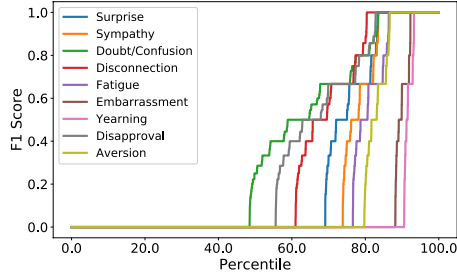
Figure 4.10: Reliability score distribution among low-performance participants (failure) and non low-performance participants (pass).

than 0.33. These distributions suggests that participants who pass annotation sanity checks and relaxed gold standard tests are more likely to be reliable. However, participants who fail at those tests may still be reliable. Therefore, conventional quality control mechanism like the gold standard is insufficient when it comes to affect data.

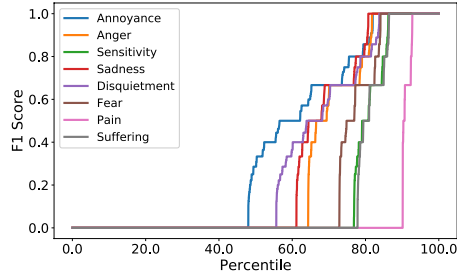
We further investigated how well humans can achieve on emotion recognition tasks. There are 5,650 AMT participants contributing to our dataset annotation. They represent over 100 countries (including 3,421 from the USA and 1,119 from India), with 48.4% male and 51.6% female, and an average age of 32. In terms of ethnicity, 57.3% self-reported as White, 21.2% Asian, 7.8% African American, 7.1% Hispanic or Latino, 1.6% American Indian or Alaskan Native, 0.4% Native Hawaiian or Other Pacific Islander, and 4.5% Other. For each participant, we used annotations from other participants and aggregated final dataset annotation to evaluate the performance. We treated this participant’s annotation as prediction from an oracle model and calculate $F1$ score for categorical emotion, and coefficient of determination (R^2) and mean squared error (MSE) for dimensional emotion to evaluate the participant’s performance. Similar to our standard annotation aggregation procedure, we ignored instances with a confidence score less than 0.95 when dealing with dimensional emotions. Fig. 4.11 shows the cumulative distribution of participants’ $F1$ scores of categorical emotions, the R^2 score, and the MSE score



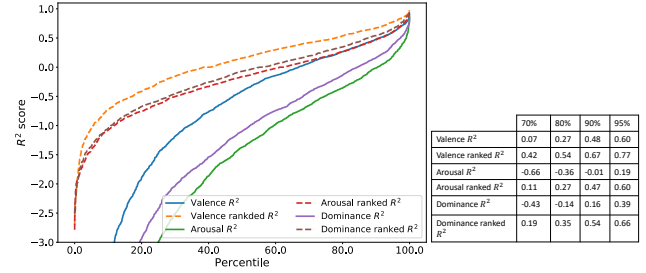
(1) $F1$ score



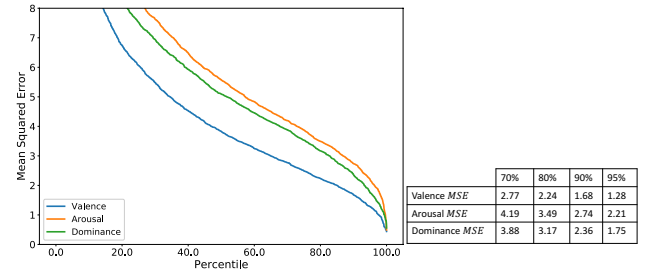
(2) $F1$ score



(3) $F1$ score



(4) R^2 score



(5) MSE

Figure 4.11: Human regression performance on dimensional emotions. X-axis: participant population percentile. Y-axis: $F1$, R^2 and MSE score. Tables inside each plot in the second row summarize top 30%, 20%, 10%, and 5% participant regression scores.

of dimensional emotion, respectively. We calculated vanilla R^2 score and rank percentile-based R^2 score. For the latter, we used rank percentile for both prediction and the ground truth. The areas under the curves (excluding Fig. 4.11(5)) can be interpreted as how difficult it is for humans to recognize the emotion. For example, humans are effective at recognizing happiness while ineffective at recognizing yearning. Similarly, humans are better at recognizing the level of valence than that of arousal or dominance. These results reflect the challenge of achieving high classification and regression performance for emotion recognition even for human beings.

4.2.4 Demographic Factors

Culture, gender, and age could be important factors of emotion understanding. As mentioned in Section 4.1.4, we have nine quality control videos in our crowdsourcing process that have been annotated for emotion more than 300 times. We used these quality control videos to test whether the annotations are independent of annotators' culture, gender, and age.

For categorical annotations (including both categorical emotions and categorical character demographics), we conducted χ^2 test on each video. For each control instance, we calculated the p-value of the χ^2 test over annotations (26 categorical emotions and 3 character demographic factors) from different groups resulting from annotators' three demographic factors. This process results in $29 \times 3 = 87$ p-value scores for each control instance. For each test among 87 pairs, we further counted the total number of videos with significant p-value ($p < 0.01$ or $p < 0.001$). Interestingly, there is significant dependence over characters' ethnicity and annotators' ethnicity (9 out of 9, $p < 0.001$). It is possible that humans are good at recognizing the ethnicity of others in the same ethnic group. Additionally, there is intermediate dependence between annotators' ethnicity and categorical emotions (17 out of $26 \times 9 = 234$, $p < 0.001$). We did not find strong dependence over other tested pairs (less than 3 out of 9, $p < 0.001$). This lack of dependence seems to suggest that a person's understanding of emotions depends more on their own ethnicity than on their age or gender.

For VAD annotation, we conducted one-way ANOVA tests on each instance. For each control instance, we calculated p-value of one-way ANOVA test over VAD (3) annotations from different groups resulting from annotators' demographic factors (3). This results in $3 \times 3 = 9$ p-value scores for each control instance. We also conducted Kruskal-Wallis H-test and found similar results. We report p-value of one-way ANOVA tests. Our results show that gender and age have little effect (less than 8 out of $9 \times (3 + 3) = 54$, $p < 0.001$) on emotion understanding, while ethnicity has a strong effect (13 out of $9 \times 3 = 27$, $p < 0.001$) on emotion understanding. Specifically, participants with different ethnicities have different understandings regarding valence for almost all control clips (7 out of 9, $p < 0.001$). Fig. 4.5(27-28) shows two control clips. For Fig. 4.5(27), valence average of person 0 among Asians is 5.56, yet 4.12 among African Americans and 4.41 among Whites. However, arousal average among Asians is 7.20, yet 8.27 among African Americans and 8.21 among Whites. For Fig. 4.5(28), valence average of person 1 among Asians is 6.30, yet 5.09 among African Americans and 4.97 among Whites. However, arousal average among Asians is 7.20, yet 8.27 among African Americans and 8.21 among Whites. Among all of our control instances, the average valence among Asians is consistently higher than among Whites and African Americans. This repeated finding seems to suggest that Asians tend to assume more positively when interpreting others' emotions.

4.2.5 Demographic Stereotype

We investigated if there are stereotypical emotionality for various demographics. According to [110], emotionality is correlated with characters’ demographics. It is worth noting that our collected annotations are not for evoked emotions from visual stimuli but for human-perceived emotional expressions of the characters. Human’s perception or assessment can be influenced by stereotypical emotionality. Based on our aggregated annotation, We conducted one-way ANOVA between each emotion dimension ($26 + 3$) and each demographical factor. For categorical emotions, there are 14 categories having significant p-value ($p < 0.001$) for gender, 10 categories for age and 13 categories for ethnicity. Specifically, women are more often depicted as affection (*frequency* = 0.18) compared with men (0.09); men are more often depicted as engagement and confidence (0.36, 0.25) compared with women (0.28, 0.15). Confidence occurs more with the increase of the age (0.09, 0.13, 0.23) and sadness occurs less with the increase of the age (0.19, 0.14, 0.08). Anticipation occurs less on African (0.17) compared with Asian (0.24) and White (0.23). Surprise occurs more on Asian (0.12) compared with African (0.07) and White (0.09). For VAD, sample mean of valence and arousal shows significant difference ($p < 0.001$) among different groups of age and ethnicity, respectively. All demographical factors show strong effect on dominance. Specifically, men are more dominating (sample mean of D: 0.61) than women (0.55); seniors are more dominating than youth (0.48, 0.53, and 0.60 for kids, teenager, and adult, respectively).

4.3 Summary

In this chapter, we describe how we created the BoLD dataset and provide results of our statistical analysis of the data. Our analysis has shown the effectiveness of the proposed data collection pipeline. Our data collection efforts offer important lessons. The efforts confirmed that reliability analysis is useful for collecting subjective annotations such as emotion labels when no gold standard ground truth is available. As shown in Table 4.1, consensus (filtered κ value) over high-reliable participants is higher than that of all participants (κ value). This finding holds for both subjective questions (categorical emotion) and objective questions (character demographics), even though the reliability score is calculated with the different VAD annotations — an evidence that the score does not overfit. As an offline quality control component, the method we developed and used to generate reliability scores [20] is suitable for analyzing such affective data. For example, one can also apply our proposed data collection pipeline to collect data for the task of image aesthetics modeling [111]. In addition to their

effectiveness in quality control, reliability scores are very useful for resource allocation. With a limited annotation budget, it is more reasonable to reward highly-reliable participants rather than less reliable ones.

Chapter 5 |

Bodily Expression Recognition

5.1 Learning from Skeleton

5.1.1 Laban Movement Analysis

Laban notation, originally proposed by Rudolf Laban ([112]), is used for documenting body movement of dancing such as ballet. Laban movement analysis (LMA) uses four components to record human body movements: body, effort, shape, and space. Body category represents structural and physical characteristics of the human body movements. It describes which body parts are moving, which parts are connected, which parts are influenced by others, and general statements about body organization. Effort category describes inherent intention of a movement. Shape describes static body shapes, the way the body interacts with something, the way the body changes toward some point in space, and the way the torso changes in shape to support movements in the rest of the body. LMA or its equivalent notation systems are widely used in psychology for emotion analysis [14, 24] and human computer interaction for emotion generation and classification [61, 74]. In our experiments, we use features listed in Table 5.1.

LMA is conventionally conducted for 3D motion capture data that have 3D coordinates of body landmarks. In our case, we estimated 2D pose on images using [27]. In particular, we denote $p_i^t \in R^2$ as the coordinate of the i -th joint at the t -th frame. As the nature of the data, our 2D pose estimation usually has missing values of joint locations and varies in scale. In our implementation, we ignored an instance if the dependencies to compute the feature are missing. To address the scaling issue, we normalized each pose by the average length of all visible limbs, such as shoulder-elbow and elbow-wrist. Let $\nu = \{(i, j) | \text{joint } i \text{ and joint } j\}$

Table 5.1: Laban Movement Analysis (LMA) features. (f_i : categories; m : number of measurements; dist.: distance; accel.: acceleration)

f_i	Description	m	f_i	Description	m
f_1	Feet-hip dist.	4	f_2	Hands-shoulder dist.	4
f_3	Hands dist.	4	f_4	Hands-head dist.	4
f_8	Centroid-pelvis dist.	4	f_9	Gait size (foot dist.)	4
f_{29}	Shoulders velocity	4	f_{32}	Elbow velocity	4
f_{13}	Hands velocity	4	f_{12}	Hip velocity	4
f_{35}	Knee velocity	4	f_{14}	Feet velocity	4
f_{38}	Angular velocity	$4C_{23}^2$			
f_{30}	Shoulders accel.	4	f_{33}	Elbow accel.	4
f_{16}	Hands accel.	4	f_{15}	Hip accel.	4
f_{36}	Knee accel.	4	f_{17}	Feet accel.	4
f_{39}	Angular accel.	$4C_{23}^2$			
f_{31}	Shoulders jerk	4	f_{34}	Elbow jerk	4
f_{40}	Hands jerk	4	f_{18}	Hip jerk	4
f_{37}	Knee jerk	4	f_{41}	Feet jerk	4
f_{19}	Volume	4	f_{20}	Volume (upper body)	4
f_{21}	Volume (lower body)	4	f_{22}	Volume (left side)	4
f_{23}	Volume (right side)	4	f_{24}	Torso height	4

j are visible} be the visible set of the instance. We computed normalized pose $\hat{p}_i^t$ by

$$s = \frac{1}{T|\nu|} \sum_{(i,j) \in \nu} \sum_t^T \|p_i^t - p_j^t\|, \quad \hat{p}_i^t = \frac{p_i^t}{s}. \quad (5.1)$$

The first part of features in LMA, *body component*, captures the pose configuration. For f_1, f_2, f_3, f_8 , and f_9 , we computed the distance between the specified joints frame by frame. For symmetric joints like feet-hip distance, we used the mean of left-feet-hip and right-feet-hip distance in each frame. The same protocol was applied to other features that contains symmetric joints like hands velocity. For f_4 , the centroid was averaged over all visible joints and pelvis is the midpoint between left hip and right hip. This feature is designed to represent barycenter deviation of the body.

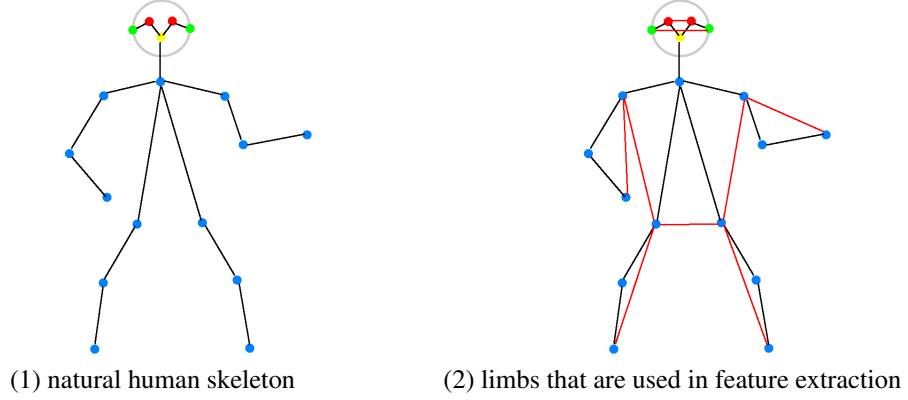


Figure 5.1: Illustration of the human skeleton. Both red lines and black lines are considered limbs in our context.

The second part of features in LMA, *effort component*, captures body motion characteristics. Based on the normalized pose, joints velocity $\hat{v}_i^t$, acceleration $\hat{a}_i^t$, and jerk $\hat{j}_i^t$ were computed as:

$$\begin{aligned} v_i^t &= \frac{\hat{p}_i^{t+\tau} - \hat{p}_i^t}{\tau}, \quad a_i^t = \frac{v_i^{t+\tau} - v_i^t}{\tau}, \quad j_i^t = \frac{a_i^{t+\tau} - a_i^t}{\tau}, \\ \hat{v}_i^t &= \|v_i^t\|, \quad \hat{a}_i^t = \|a_i^t\|, \quad \hat{j}_i^t = \|j_i^t\|. \end{aligned} \quad (5.2)$$

Angles, angular velocity, and angular acceleration between each pair of limbs (Fig. 5.1) were calculated for each pose:

$$\begin{aligned} \theta^t(i, j, m, n) &= \arccos \left(\frac{(\hat{p}_i^t - \hat{p}_j^t) \cdot (\hat{p}_m^t - \hat{p}_n^t)}{\|\hat{p}_i^t - \hat{p}_j^t\| \|\hat{p}_m^t - \hat{p}_n^t\|} \right), \\ \omega_k^t(i, j, m, n) &= \frac{\theta^{t+\tau}(i, j, m, n) - \theta^t(i, j, m, n)}{\tau}, \\ \alpha_k^t(i, j, m, n) &= \frac{\omega^{t+\tau}(i, j, m, n) - \omega^t(i, j, m, n)}{\tau}. \end{aligned} \quad (5.3)$$

We computed velocity, acceleration, jerk, angular velocity, and angular acceleration of joints with $\tau = 15$. Empirically, features become less effective when τ is too small ($1 \sim 2$) or too large (> 30).

The third part of features in LMA, *shape component*, captures body shape. For f_{19} , f_{20} , f_{21} , f_{22} , and f_{23} , the area of bounding box that contains corresponding joints is used to approximate volume.

Finally, all features are summarized by their basic statistics (maximum, minimum, mean, and standard deviation, denoted as $f_i^{\max}$, $f_i^{\min}$, f_i^{mean} , and f_i^{std} , respectively) over time.

With all LMA features combined, each skeleton sequence can be represented by a 2,216-D feature vector. We further build classification and regression models for bodily expression recognition tasks. Because some measurements in our feature set can be linearly correlated and features can be missing, we choose the random forest for our classification and regression task. Specifically, we impute missing feature values with a large number (1,000 in our case). We then search model parameters with cross validation on the combined set of training and validation. Finally, we use the selected best parameter to retrain a model on the combined set.

5.1.2 Spatial Temporal Graph Convolutional Network

Besides handcrafted LMA features, we experimented with an end-to-end feature learning method. Following [5], human body landmarks can be constructed as a graph with their natural connectivity. Considering the time dimension, a skeleton sequence could be represented with a spatiotemporal graph. Graph convolution in [113] is used as building blocks in ST-GCN. ST-GCN was originally proposed for skeleton action recognition. In our task, each skeleton sequence is first normalized between 0 and 1 with the largest bounding box of skeleton sequence. Missing joints are filled with zeros. We used the same architecture as in [5] and trained on our task with binary cross-entropy loss and mean-squared-error loss. Our learning objective $\mathcal{L}$ can be written as:

$$\begin{aligned}\mathcal{L}_{\text{cat}} &= \sum_{i=1}^{26} y_i^{\text{cat}} \log x_i + (1 - y_i^{\text{cat}}) \log(1 - x_i^{\text{cat}}) , \\ \mathcal{L}_{\text{cont}} &= \sum_{i=1}^3 (y_i^{\text{cont}} - x_i^{\text{cont}})^2 , \\ \mathcal{L} &= \mathcal{L}_{\text{cat}} + \mathcal{L}_{\text{cont}} ,\end{aligned}\tag{5.4}$$

where x_i^{cat} and y_i^{cat} are predicted probability and ground truth, respectively, for the i -th categorical emotion, and x_i^{cont} and y_i^{cont} are model prediction and ground truth, respectively, for the i -th dimensional emotion.

5.2 Learning from Pixels

Essentially, bodily expression is expressed through body activities. Activity recognition is a popular task in computer vision. The goal is to classify human activities, like sports and housework, from videos. In this subsection, we use four classical human activity recognition methods to extract features [6–9]. Current state-of-the-art results of activity recognition are achieved by two-stream network-based deep-learning methods [7]. Prior to that, trajectory-

based handcrafted features are shown to be efficient and robust [114, 115].

5.2.1 Trajectory based Handcrafted Features

The main idea of trajectory-based feature extraction is selecting extended image features along point trajectories. Motion-based descriptors, such as histogram of flow (HOF) and motion boundary histograms (MBH) [116], are widely used in activity recognition for their good performance [114, 115]. Common trajectory-based activity recognition has the following steps: 1) computing the dense trajectories based on optical flow; 2) extracting descriptors along those dense trajectories; 3) encoding dense descriptors by Fisher vector [117]; and 4) training a classifier with the encoded histogram-based features.

In this work, we cropped each instance from raw clips with a fixed bounding box that bounds the character over time. We used the implementation in [6] to extract trajectory-based activity features¹. We trained 26 SVM classifiers for the binary categorical emotion classification and three SVM regressors for the dimensional emotion regression. We selected the penalty parameter based on the validation set and report results on the test set.

5.2.2 Deep Activity Features

Two-stream network-based deep-learning methods learn to extract features in an end-to-end fashion [7]. A typical model of this type contains two convolutional neural networks (CNN). One takes static images as input and the other takes stacked optical flow as input. The final prediction is an averaged ensemble of the two networks. In our task, we used the same learning objective of $\mathcal{L}$ as defined in Eq. 5.4.

We implemented two-stream networks in PyTorch². We used 101-layer ResNet as [118] as our network architecture. Optical flow was computed via TVL1 optical flow algorithm [119]. Both image and optical flow were cropped with the instance body centered. Since emotion understanding could be potentially related to color, angle, and position, we did not apply any data augmentation strategies. The training procedure is identical to the work of [7], where the learning rate is set to 0.01. We used resnet-101 model pretrained on ImageNet to initialize our network weights. The training takes around 8 minutes for one epoch with an NVIDIA Tesla K40 card. The training time is short because only one frame is sampled input for each video in the RGB stream, and 10 frames are concatenated along the channel dimension in the optical flow stream. We used the validation set to choose the model of the lowest loss. We name this

¹<https://github.com/vadimkantorov/fastvideofeat>

²<http://pytorch.org/>

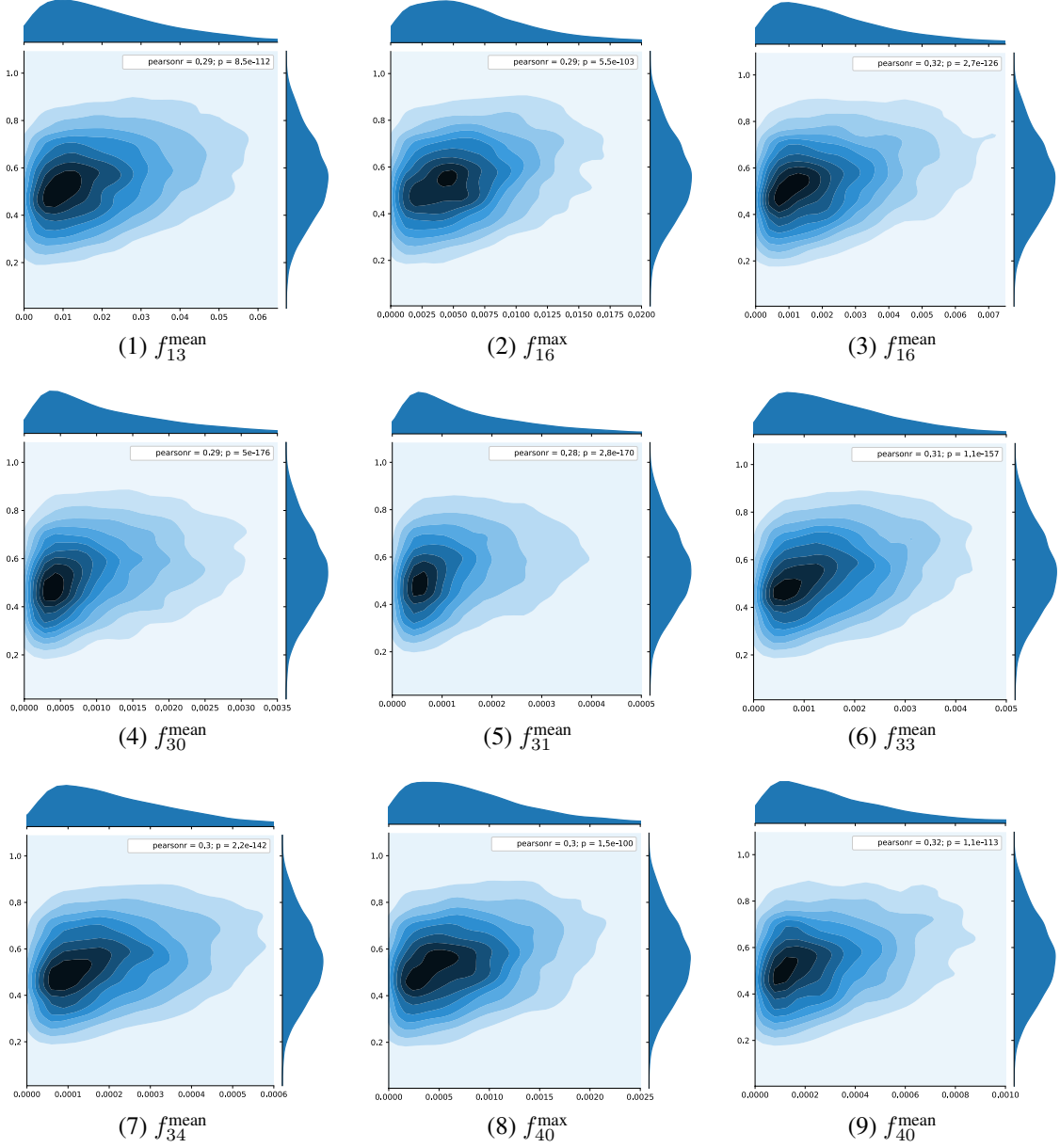


Figure 5.2: Kernel density estimation plots on selected LMA features that have high correlation with arousal.

model as TS-ResNet101.

Besides the original two-stream network, we also evaluated its two other state-of-the-art variants of action recognition. For temporal segment networks (TSN) [9], each video is divided into K segments. One frame is randomly sampled for each segment during the training stage. Video classification result is averaged over all sampled frames. In our task, learning rate is set to 0.001 and batch size is set to 128. For two-stream inflated 3D ConvNet (I3D) [8], 3D

convolution replaces 2D convolution in the original two-stream network. With 3D convolution, the architecture can learn spatiotemporal features in an end-to-end fashion. This architecture also leverages recent advances in image classification by duplicating weights of pretrained image classification model over the temporal dimension and using them as initialization. In our task, learning rate is set to 0.01 and batch size is set to 12. Both experiments are conducted on a server with two NVIDIA Tesla K40 cards. Other training details are the same as the original work [8, 9].

5.3 Results

5.3.1 Evaluation Metrics

We evaluated all methods on the test set. For categorical emotion, we used average precision (AP, area under precision recall curve) and area under receiver operating characteristic curve (ROC AUC) to evaluate the classification performance. For dimensional emotion, we used R^2 to evaluate regression performance. Specifically, a random baseline of AP is the proportion of the positive samples (P.P.). ROC AUC could be interpreted as the possibility of choosing the correct positive sample among one positive sample and one negative sample; a random baseline for that is 0.5. To compare performance of different models, we also report mean R^2 score (mR^2) over three dimensional emotion, mean average precision (mAP), and mean ROC AUC (mRA) over 26 categories of emotion. For the ease of comparison, we define emotion recognition score (ERS) as follows and use it to compare performance of different methods:

$$\text{ERS} = \frac{1}{2} \left(mR^2 + \frac{1}{2}(\text{mAP} + \text{mRA}) \right) . \quad (5.5)$$

5.3.2 LMA Feature Significance Test

For each categorical emotion and dimension of VAD, we conducted linear regression tests on each dimension of features listed in Table 5.1. All tests were conducted using the BoLD training set. We did not find strong correlations ($R^2 < 0.02$) over LMA features and emotion dimensions other than arousal, *i.e.*, categorical emotion and valence and dominance. Arousal, however, seems to be significantly correlated with LMA features. Fig. 5.2 shows the kernel density estimation plots of features with top R^2 on arousal. Hands-related features are good indicators for arousal. With hand acceleration, f_{16}^{mean} alone, R^2 can be achieved as 0.101. Other significant features for predicting arousal are hands velocity, shoulders acceleration, elbow acceleration, and hands jerk.

Table 5.2: Dimensional emotion regression and categorical emotion classification performance on the test set. mR^2 = mean of R^2 over dimensional emotions, $mAP(\%)$ = average precision / area under precision recall curve (PR AUC) over categorical emotions, $mRA(\%)$ = mean of area under ROC curve (ROC AUC) over categorical emotions, and ERS = emotion recognition score. Baseline methods: ST-GCN [5], TF [6], TS-ResNet101 [7], I3D [8], and TSN [9].

Model	Regression	Classification		ERS
	mR^2	mAP	mRA	

A Random Method based on Priors:

Chance	0	10.55	50	0.151
--------	---	-------	----	-------

Learning from Skeleton:

ST-GCN	0.044	12.63	55.96	0.194
LMA	0.075	13.59	57.71	0.216

Learning from Pixels:

TF	-0.008	10.93	50.25	0.149
TS-ResNet101	0.084	17.04	62.29	0.240
I3D	0.098	15.37	61.24	0.241
TSN	0.095	17.02	62.70	0.247
TSN-Spatial	0.048	15.34	60.03	0.212
TSN-Flow	0.098	15.78	61.28	0.241

5.3.3 Model Performance

Table 5.2 shows the results on the emotion classification and regression tasks. TSN achieves the best performance, with a mean R^2 of 0.095, a mean average precision of 17.02%, a mean ROC AUC of 62.70%, and an ERS of 0.247. Fig. 5.3 presents detailed metric comparisons over all methods of each categorical and dimensional emotion.

For the pipeline that learns from the skeleton, both LMA and ST-GCN achieved above-chance results. Our handcrafted LMA features performs better than end-to-end ST-GCN under all evaluation metrics. For the pipeline that learns from pixels, trajectory-based activity features did not achieve above-chance results for both regression and classification task. However, two-stream network-based methods achieved significant above-chance results for both regression and classification tasks. As shown in Fig. 5.3 and Table 4.1, most top-performance categories, such as affection, happiness, pleasure, excitement, sadness, anger, and pain, receive high

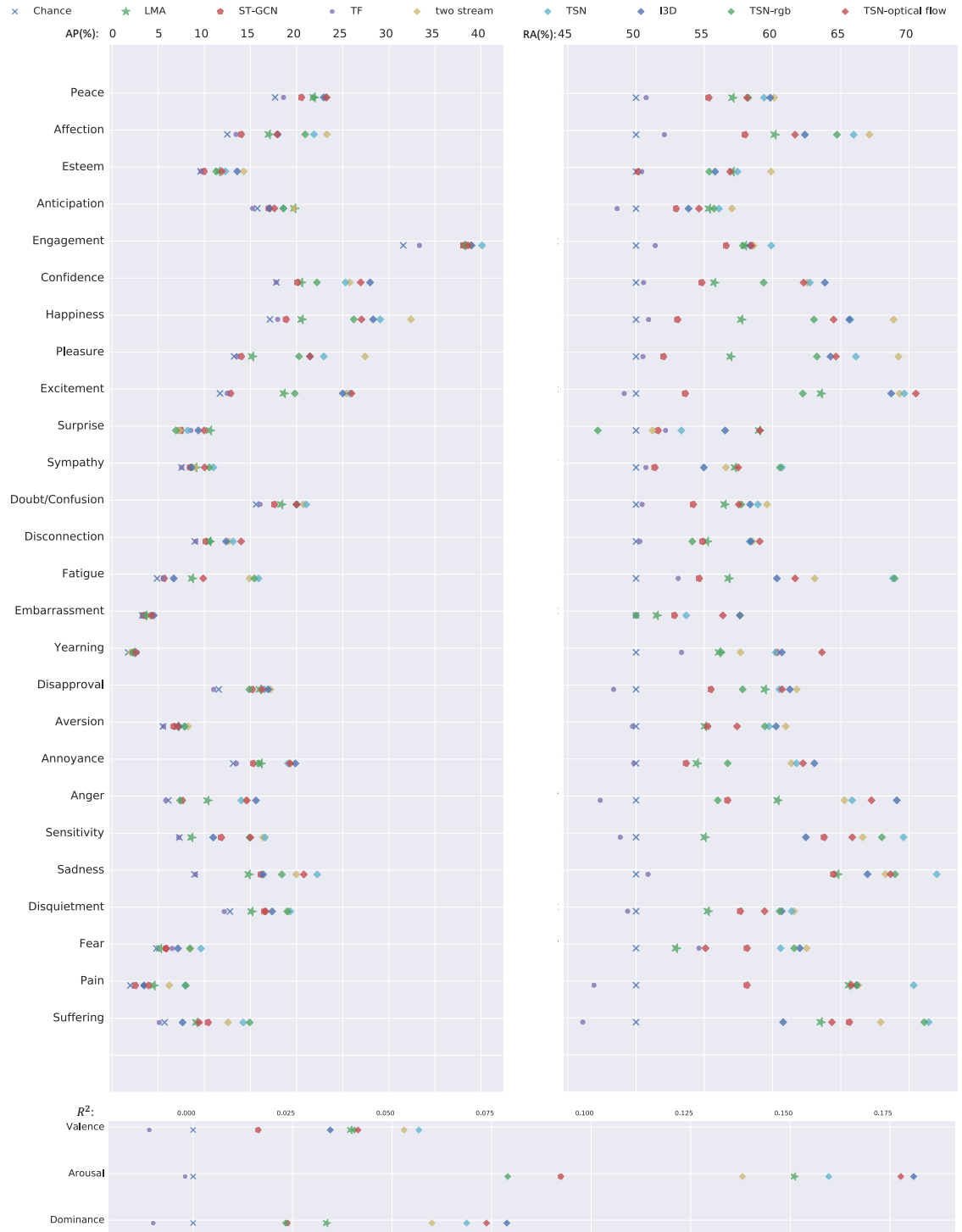


Figure 5.3: Classification performance (AP: average precision on the top left, RA: ROC AUC on the top right) and regression performance (R^2 on the bottom) of different methods on each categorical and dimensional emotion.

agreement (κ) among annotators. Similar to the results from skeleton-based methods, two-stream network-based methods show better regression performance over arousal than for valence and dominance. However, as shown in Fig. 4.11, workers with top 10% performance has R^2 score of 0.48, -0.01 , and 0.16 for valence, arousal, and dominance, respectively. Apparently, humans are best at recognizing valence and worst at recognizing arousal, and the distinction between human performance and model performance may suggest that there could be other useful features that the model has not explored.

5.3.4 Ablation Study

To further understand the effectiveness of the two-stream-based model on our task, we conducted two sets of experiments to diagnose 1) if our task could leverage learned filters from pretrained activity-recognition model, and 2) how much a person’s face contributed to the performance in the model. Since TSN has shown the best performance among all two-stream-based models, we conducted all experiments with TSN in this subsection. For the first set of experiments, we used different pretrained models, *i.e.*, image-classification model pretrained on ImageNet [120] and action recognition model pretrained on Kinetics [65], to initialize TSN. Table 5.3 shows the results for each case. The results demonstrate that initializing with pretrained ImageNet model leads to slightly better emotion-recognition performance. For the second set of experiments, we train TSN with two other different input types, *i.e.*, face only and faceless body. Our experiment in the last section crops the whole human body as the input. For face only, we crop the face for both spatial branch (RGB image) and temporal branch (optical flow) during both the training and testing stages. Note that for the face-only setting, orientation of faces in our dataset may be inconsistent, *i.e.* facing forward, facing backward, or facing to the side. For the faceless body, we still crop the whole body, but we also mask the region of face by imputing pixel value with a constant 128. Table 5.4 shows the results for each setting. We can see from the results that the performance of using either the face or the faceless body as input is comparable to that of using the whole body as input. This result suggests both face and the rest of the body contribute significantly to the final prediction. Although the “whole body” setting of TSN performs better than any of the single model do, it does so by leveraging both facial expression and bodily expression.

Table 5.3: Ablation study on the effect of pretrained models.

Pretrained Model	Regression	Classification		ERS
	mR^2	mAP	mRA	
ImageNet	0.095	17.02	62.70	0.247
Kinetics	0.093	16.77	62.53	0.245

Table 5.4: Ablation study on the effect of face.

Input Type	Regression	Classification		ERS
	mR^2	mAP	mRA	
whole body	0.095	17.02	62.70	0.247
face only	0.092	16.21	62.18	0.242
faceless body	0.088	16.61	62.30	0.241

Table 5.5: Ensembled results.

Model	Regression	Classification		ERS
	mR^2	mAP	mRA	
TSN-body	0.095	17.02	62.70	0.247
TSN-body + LMA	0.101	16.70	62.75	0.249
TSN-body + TSN-face	0.101	17.31	63.46	0.252
TSN-body + TSN-face + LMA	0.103	17.14	63.52	0.253

5.3.5 ARBEE: Automated Recognition of Bodily Expression of Emotion

We constructed our emotion recognition system, ARBEE, by ensembling best models of different modalities. As suggested in the previous subsection, different modalities could provide complementary clues for emotion recognition. Concretely, we average the prediction from different models (TSN-body: TSN trained with whole body, TSN-face: TSN trained with face, and LMA: random forest model with LMA features) and evaluate the performance on the test set. Table 5.5 shows the results of ensembled results. According to the table, combining all modalities, *i.e.*, body, face and skeleton, achieves the best performance. ARBEE is the average

Table 5.6: Retrieval results of our deep model. $P@K(\%)$ = precision at K, $R-P(\%)$ =R-Precision.

Category	P@10	P@100	R-P
Peace	40	33	28
Affection	50	32	26
Esteem	30	14	12
Anticipation	30	24	20
Engagement	50	46	42
Confidence	40	33	31
Happiness	30	36	31
Pleasure	40	25	23
Excitement	50	41	31
Surprise	20	6	8
Sympathy	10	14	12
Doubt/Confusion	20	33	25
Disconnection	20	20	18
Fatigue	40	20	17
Embarrassment	0	5	5
Yearning	0	2	4
Disapproval	30	28	22
Aversion	10	10	11
Annoyance	30	28	23
Anger	40	24	20
Sensitivity	30	19	19
Sadness	50	34	25
Disquietment	10	26	25
Fear	10	8	8
Pain	20	9	12
Suffering	10	17	18
Average	27	23	20

ensemble of the three models.

We further investigated how well ARBEE retrieves instances in the test set given a specific categorical emotion as query. Concretely, we calculated precision at 10, 100, and R-Precision as summarized in Table 5.6. R-Precision is computed as precision at R , where R is number of positive samples. Similar to the classification results, happiness and pleasure can be retrieved with a rather high level of precision.

5.4 Summary

In this chapter, we investigated two pipelines for automated recognition of bodily expression and present quantitative results for some baseline methods. Unlike AMT participants, who were provided with all the information regardless of whether they use all in their annotation process, the first computerized pipeline relied solely on body movements, but *not* on facial expressions, audio, or context. The second pipeline took a sequence of cropped images of the human body as input, without explicitly modeling facial expressions. We built effective predictive models for bodily expression recognition with both pipelines.

Chapter 6 |

Learning from Depth: Image-based 3D Human Mesh Reconstruction using Unlabeled RGB-D Data

6.1 Introduction

The visual sensors used in common cameras fold the 3-dimensional (3D) world into 2-dimensional (2D) data. The gap caused by this missing dimension creates fundamental challenges in 3D computer vision. In this chapter, we focus on one of those problems—generating 3D human mesh reconstruction from a single 2D RGB photo. Recognizing human behaviors and detecting emotional or mental states are of foundational importance to machine perception since many applications depend on them [101, 121]. In the past few years, the capability of machine perception evolved from generating human bounding boxes to detecting human key points, and further to producing comprehensive representations. A fine-grained mesh of human pose and shape can provide much richer geometric information than articulated joints alone. Yet, accurately estimating mesh to faithfully represent a human in the view is highly challenging for machine learning. Because there is no ground-truth representation of meshes in the physical world, existing works have formulated the task as a machine learning problem, where the main technical obstacle is to obtain the mesh information from a proper source. One common approach is to project human joints on a 3D mesh onto the 2D domain, and to acquire their 2D ground-truth locations as the supervision signal in model training [97]. Here, supervision signal means the response variables in the training data, the counterpart of

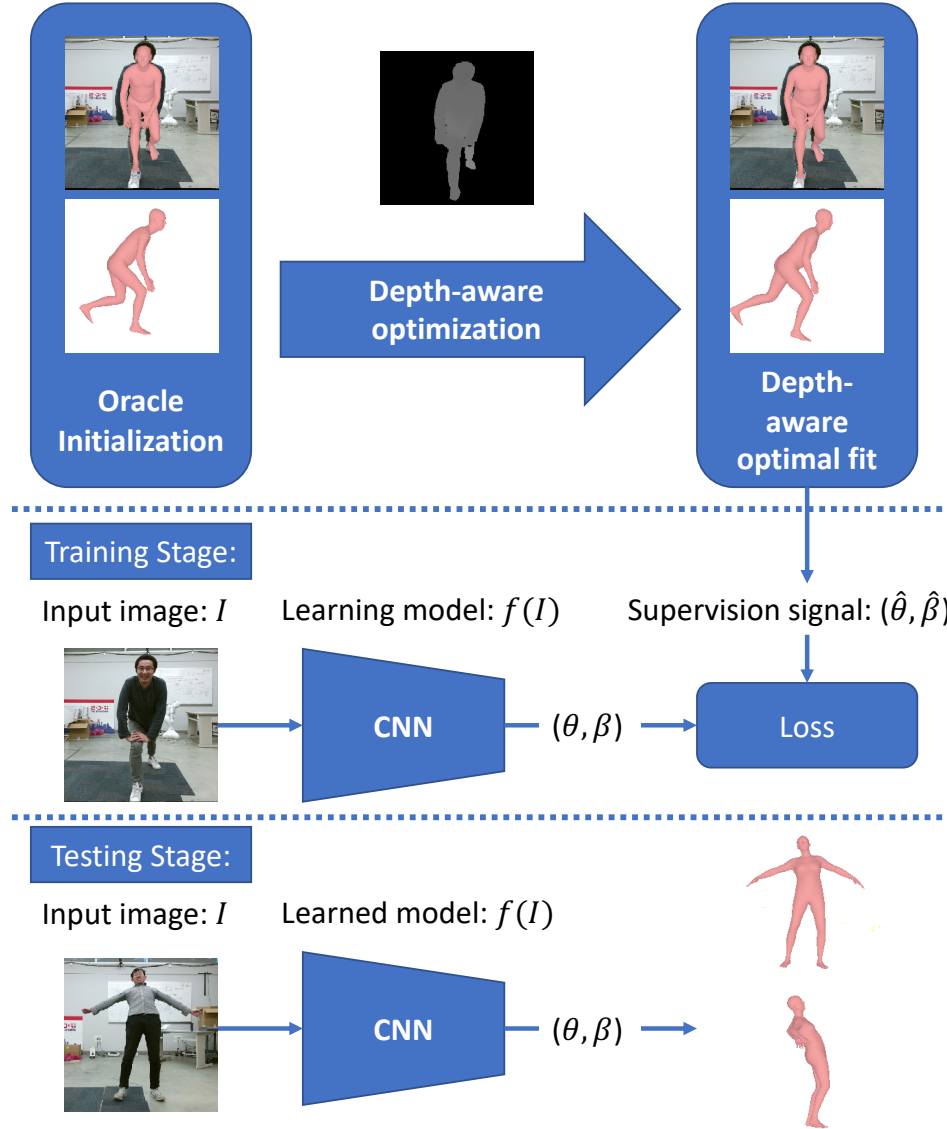


Figure 6.1: Our method optimizes the human oracle from an image-based initialization to one being also consistent with additional depth data. The resulting optimized human model parameters are then used as supervision for a machine learning system that only takes a 2D image as input and generates human pose and shape, *i.e.*, parameters of the SMPL parametric model [4].

ground-truth labels in a classification problem. Those 2D ground-truth locations are often generated by a manual labeling process. Another approach is to directly train a model to fit 3D human joints, where those joint labels can be generated by an indoor motion capture (MoCap) system [97, 122, 123]. Neither approach is cost-effective or readily scalable for real-world applications. When one needs to deploy a learning model for a task carried out in a wildly different

environment or with data collection from non-standard sensors, they inevitably acquire sensor or environment-specific data labels from scratch. Moreover, neither of the aforementioned types of supervising signals captures fully the information of the human shape. 3D/2D joints could indicate the limb/torso length and thus how tall or short a person is. However, they cannot suggest the thickness of limb/torso and thus the bulkiness of a body.

RGB-D images provide a complementary data source for training a 3D human mesh reconstruction model from a single RGB image. RGB-D data are widely available in the existing literature and convenient to collect if desired. As a common practice, modern semi-supervised learning methods can reap the benefits of additional unlabeled data in the same form as the input to the learning system (*i.e.*, the RGB images in our case), but is not designed to handle additional information (such as the depth). We argue that depth is particularly valuable and can naturally serve as a supervision signal for the task, scaling up learning models to a level not limited by the high cost of acquiring labels. As a result, not surprisingly, the mesh reconstruction model trained with those additional RGB-D data can be substantially more accurate than traditional semi-supervised approaches which do not exploit the additional depth dimension.

In this chapter, we explore the intuitive idea of “learning from depth” in two ways: (1) We develop a method easy for incorporation into the existing human mesh reconstruction framework to gain “depth” knowledge from RGB-D training data. (2) We evaluate the model accuracy from the depth perspective in addition to other conventional metrics to demonstrate the complementary benefits of semi-supervised learning and “learning from depth”.

6.2 Related Work

Learning from Depth: Learning depth information from monocular images is a well-explored topic in computer vision. Although learning depth from monocular images is not by the approach of geometric triangulation, it has been shown empirically a depth estimation model can be helpful [124, 125]. There has been interesting analysis on how a neural network understands depth from an outdoor-scene monocular image [126]. Specifically, to estimate human pose and shape, both prior knowledge and detailed textures are important sources of information, especially in the depth dimension. For example, one can interpret the correct depth for an elbow based on the fact that an elbow cannot bend towards the other side. Most approaches relies heavily on such prior information. For instance, [127] learns such prior knowledge from 3D human pose scan data and penalizes unrealistic poses accordingly. [97] proposes to constrain the predicted human pose to the prior manifold by adding an adversarial

prior during the training process. However, most existing approaches do not make extensive use of the texture information from a monocular image. Specifically, the height and bulkiness of a person can be observed directly from a monocular image. The height of a child and an adult is significantly different, which can be inferred from the image [128]. The reason for not fully leveraging texture information is that the ground truth fails to capture enough details at the first place, even though human model like SMPL is capable of representing such details. Therefore, we believe that our proposed approach to exploit detailed depth information is an important step towards enriching the representation of a human model.

Semi-supervised Learning: Applying a pre-trained model to generate pseudo labels for unlabeled instances has been a popular approach of semi-supervised learning [129–132]. This approach can potentially reduce a model’s variance, yet, it is unlikely to overcome bias in the model trained on a small labeled data set. A main technical issue is that incorrect pseudo labels can confirm bias in training. There are attempts to mitigate the issue by estimating uncertainty for pseudo labels [133, 134]. We argue that we can further mitigate the problem from a practical data collection perspective. In our case (a.k.a., training human mesh predictions), we can significantly improve pseudo labels’ quality by leveraging additional depth data.

Human Mesh Reconstruction: Detecting and representing humans in images have been an important area in computer vision. The representation of humans has evolved from coarse-grained ones, like bounding boxes [135], 2D poses [136] and 3D poses [123], to fine-grained ones, like mesh based parametric human models [4, 137, 138] and implicit function based representation [139]. As the level of granularity grows from coarse to fine, the difficulty of acquiring ground truth increases significantly, if possible at all. Annotating 2D bounding boxes and 2D poses from images mostly requires manual efforts, scalable with the modern crowd-sourcing frameworks. However, fine-grained human representations are not feasible for manual annotation. For example, generating ground truth for SMPL human model representation requires the setup of MoCap system and putting multiple markers on the subject for accurate estimation [10]. Other work [127] that only leverages 2D pose or sparse 3D pose to recover human mesh reconstruction must trade off between the accuracy of the pose recoverability and the feasibility of data collection. Given those challenges in acquiring high quality fine granularity ground-truth, a recent work [140] proposes to generate 3D human mesh with a state-of-the-art method and ask human annotators to select good fits such that a large dataset of 3D body models can be curated. Another work [141] proposes to use a neural network to generate pseudo labels for in-the-wild images. [12] provides an interesting paradigm that incorporates the training process and the ground truth generation in one loop. That work is based on the fact that optimization-based ground truth generation methods, e.g., [127],

rely mostly on a good initialization at the beginning of pose configuration. [142] proposes to improve the quality of ground truth by performing bundle adjustment on a video sequence of a single person. [143] proposes a novel setup for collecting 3D human poses with multiple IMUs attached to the limb and human shape with 3D scanner. [144] proposes to manually annotate the ordinal depth relationship of two joints for 3D pose estimation.

Different from previous approaches, we propose to improve ground truth of human pose and shape at the stage of raw data collection: acquiring depth image at the same time that RGB image is captured. Such a setup is cost effective as more and more mobile devices are equipped with short-ranged depth sensors.

6.3 Technical Approach

We first review the parametric human body model SMPL [4] and define the notations in Sec. 6.3.1. Next, we describe our proposed approach for generating accurate optimized SMPL parameters when additional depth data are available in Sec. 6.3.2. Those optimized parameters suppose to serve as the pseudo ground-truth in learning. Lastly, we introduce the baseline learning approach used in our experiments in Sec. 6.3.3. The learning framework consumes pseudo ground-truth parameters introduced in Sec.6.3.2 and predicts more accurate human meshes.

6.3.1 The SMPL Model

The SMPL body model [4] is a parametric model that takes two sets of parameters, the pose parameters θ and the shape parameters β , as inputs and produces human body mesh $\mathbf{M}(\theta, \beta) \in \mathbb{R}^{N \times 3}$, where it is spanned by $N = 6890$ vertices $[v_1, \dots, v_N]^T$. The pose parameters θ are an axis-angle representation of the relative rotation $SO(3)$ for each articulated joint on the human body. For the SMPL model, there are $k = 23$ human body joints $\{x_1, \dots, x_k\}$. Their 3D position $\mathbf{J}$ can be further acquired with a fixed sparse linear combination of the vertices $\mathbf{M}$, *i.e.*,

$$\mathbf{J}(\theta, \beta) = \mathbf{W}\mathbf{M}(\theta, \beta) \in \mathbb{R}^{k \times 3},$$

where $\mathbf{W}$ is pre-determined coefficients from real human mesh data collected by a 3D laser scanner. The shape parameters β are essentially weights of orthonormal principal components of shape displacements. The output human mesh can represent different body types by adjusting the shape parameters. The SMPL body model generates human mesh by rotating vertices group over corresponding joints from a ‘T’-pose human mesh, which is originally crafted by artist and then morphed with linear terms to rectify the impact of target pose (θ) and shape (β).

6.3.2 Depth-aware Ground Truth Generation

Estimating accurate parameters (θ, β) of the SMPL model for a posed human is an optimization problem. The objective is minimizing the displacement between model-rendered human mesh $\mathbf{M}$ and real observations from sensors $\{o_k\}$, *i.e.*,

$$\theta^*, \beta^* = \arg \min_{\theta, \beta} \sum_k \rho(F(\mathbf{M}(\theta, \beta), k), o_k), \quad (6.1)$$

where ρ can be either simply L2 loss or robust estimators like the Geman-McClure penalty function [145], depending on the fidelity of the observation sensor; $F(\cdot)$ is a function that corresponds k -th observations with mesh vertices. In particular, if o_k is the 3D position of body joints and measured by MoCap system with a physical marker on body joint k , F is the k -th joint x_k , *i.e.*, $F(\mathbf{M}, k) = (\mathbf{WM})_k$. In more general cases when observations are full point cloud and measured by the 3D scanner, one can first register the observed point cloud on the model-rendered mesh and then sort out the correspondence [146]. Thus, $F(\mathbf{M}, k)$ can be the vertices v_t or a linear combination of a set of close vertices $(v_{t_1}, \dots, v_{t_n})$ based on the registered correspondence.

In our case where depth is measured, the observation “depth” is essentially either a dense point cloud from a time-of-flight depth sensor or a sparse point cloud from a Lidar-based point cloud. Without loss of generality, we assume the depth observation can be eventually transformed back to the RGB camera’s coordinate system through calibration. We denote the resulting depth image as I_d . Note that I_d can contain missing values due to sensor limitation. For any given camera intrinsics K and six degree of freedom (6DoF) transformation P between SMPL model’s and RGB camera’s coordinate systems, we can project any point X on the mesh M to the image plane with weak perspective projection: $x = KPX$. Note that X has the homogeneous coordinates form. Due to self-occlusion, many 3D points can eventually land on the same position on the 2D image plane. Keeping the points that are closest to the camera for the same position on the 2D image plane will result in the rendered depth map, I_{rd} . In our implementation, we used a neural 3D mesh renderer [147] so that the gradient can be back-propagated to the parameters. Following [12, 127], we set fixed camera intrinsics K and estimate camera pose P by solving a perspective-n-point problem upon four pairs of 3D-2D correspondence, *i.e.*, left/right shoulder and left/right hip.

Up to this point, we have both model-rendered depth image I_{rd} and observed depth image I_d , which correspond to $\mathbf{M}$ and $\{o_k\}$ in Eq.(4.3), respectively. Empirically, we find it is efficient enough to assume exact point-wise correspondence between I_d and I_{rd} . Therefore, the objective

in Eq.(6.1) can be written as the following in our case:

$$\begin{aligned}\hat{I}_d &= I_d - I_d^{\text{pelvis}}, \quad \hat{I}_{rd}(\theta, \beta) = I_{rd}(\theta, \beta) - I_{rd}^{\text{pelvis}}(\theta, \beta), \\ E_{\text{depth}}(\theta, \beta) &= \frac{\sum_{\mathbf{p}} \delta^{\mathbf{p}} \left(\hat{I}_d^{\mathbf{p}} - \hat{I}_{rd}^{\mathbf{p}}(\theta, \beta) \right)^2}{\sum_{\mathbf{p}} \delta^{\mathbf{p}}},\end{aligned}\tag{6.2}$$

where $\mathbf{p}$ is the position in the 2D image plane, $\delta^{\mathbf{p}}$ is 1 if the depth values on both I_d and I_{rd} are not missing, and I_d^{pelvis} and I_{rd}^{pelvis} are depth values at the position of pelvis for measured and rendered depth image respectively.

While an observed depth map provides detailed depth information of a pose, it does not necessarily correspond to a unique pose. When the 2D pose is available, the pose is more likely to be uniquely determined. Meanwhile, prior assumptions, *e.g.*, elbow and knee cannot bend towards the other side, can be useful for preventing unrealistic poses. Therefore, we incorporate the above objective with the simplified formulation of SMPLify [12, 127]. Our final optimization objective can be written as:

$$\begin{aligned}E_J(\beta, \theta; K, J_{\text{est}}) + \lambda_{\theta} E_{\theta}(\theta) + \lambda_a E_a(\theta) + \\ \lambda_{\beta} E_{\beta}(\beta) + \lambda_d E_{\text{depth}}(\theta, \beta),\end{aligned}\tag{6.3}$$

where J_{est} are 2D joints' positions on the image plane that are either estimated by a convolutional neural network (CNN) model or annotated as ground truth, and K are the camera intrinsics. The first term E_J is a reprojection loss between a 2D pose oracle J_{est} and the projected SMPL joints. $E_{\theta}(\theta)$ is the log likelihood of a mixture of Gaussians pose prior trained on real-world poses. $E_a(\theta)$ is an exponential loss that penalizes elbow and knee hyperextending. $E_{\beta}(\beta)$ is a quadratic penalty on the shape coefficients. For details about the other error terms in Eq. 6.3, we refer the reader to [127]. Note that we followed [12] and did not include the interpenetration error term in the original SMPLify formulation [127].

6.3.3 Learning Solution

Given a large number of RGB-D images, it is expected that the estimated ground truth with the above optimization formulation would incubate a more accurate data-driven model that directly reconstructs human poses and shape from a single image. In our work, we use a baseline deep CNN [12, 97] to validate the effectiveness of our ground-truth generation approach. The network $f(\cdot)$ takes an RGB image I as input and directly regresses parameters of the SMPL model as well as the camera extrinsics as output, *i.e.*, $(\theta, \beta, P) = f(I)$. Specifically, the backbone of the

network is ResNet50 [118]. Reprojection loss is commonly used for supervision:

$$L_{2D} = \left\| J_{\text{pred}}^{2D} - J_{\text{gt}}^{2D} \right\|^2, \quad (6.4)$$

where J_{pred}^{2D} is the estimated 2D pose based on the network outputs (θ, β, P) . J_{gt}^{2D} are the ground truth of the 2D pose. When the ground truth of 3D pose or SMPL model parameters are available, we can also directly apply supervision on them:

$$L_{3D} = \left\| J_{\text{pred}}^{3D} - J_{\text{gt}}^{3D} \right\|^2, \quad (6.5)$$

$$L_{\text{SMPL}} = \left\| \theta_{\text{pred}} - \theta_{\text{gt}} \right\|^2 + \left\| \beta_{\text{pred}} - \beta_{\text{gt}} \right\|^2. \quad (6.6)$$

As mentioned in [12], directly supervising over SMPL parameters is essentially reducing the search space for the network due to the ambiguity of the 2D pose itself. Moreover, SMPL parameters form a richer representation of a human pose and shape than 2D joints. By directly supervising over SMPL parameters, the supervision signal is providing more guidance on human shape instead of the pose alone from 2D/3D joints.

6.4 Experiments

6.4.1 Training Datasets

We applied our ground-truth generation approach on the NTU RGB+D 120 dataset [148], which is a large RGB-D video dataset for human action recognition collected with Microsoft Kinect V2 cameras. The dataset contains 114,480 videos from 106 subjects. Human actions like hand waving are recorded in a lab setting environment with three fixed-position Kinect devices covering different perspectives simultaneously. There are 32 different setups of camera’s height (ranging from 0.5m to 2.7m) and distance (ranging from 2m to 4.5m) to the subject. We choose 52 among 120 different actions that do not involve interactions with objects or other human subjects, such as “putting on jacket” or “hugging”, because otherwise we will need to distinguish object or other human depth from human depth. To avoid duplicated training samples of the same image, we further downsampled all video frames by a fifth. The original dataset did not provide calibration files between depth camera and the RGB camera; and the same pair of an RGB image and a depth image is not pixel-wise aligned. However, human joints predictions from the Kinect API on both the RGB image and depth image are available. We fitted a projective transformation based on two set of 2D joints positions from

Datasets	# samples	2D	3D	SMPL
Human3.6M [123]	312, 188	✓	✓	✓* [10]
MPI-INF-3DHP [150]	96, 507	✓	✓	✓* [127]
LSP [151]	1, 000	✓	✗	✓* [127]
LSP-Extended [152]	9, 428	✓	✗	✓* [127]
MPII [136]	14, 810	✓	✗	✓* [127]
COCO [153]	28, 344	✓	✗	✓* [127]
NTU RGB+D [148]	680, 932	✓*	✗	✓* (ours)
HDE [154]	22, 233	✓*	✗	✓* (ours)

Table 6.1: Various datasets used and their availability of different supervision signals: 2D pose, 3D pose, and SMPL parameters. ✓ means the supervision signal is manually annotated or directly recorded from the sensor. ✓* indicates the supervision signal being estimated from an oracle approach. ✗ indicates the supervision signal being unavailable. Note that SMPL parameters cannot be either manually annotated or directly recorded due to the nature of the SMPL parametric model. We refer readers to the corresponding references on how SMPL parameters were estimated. We remark that MoSh [10] is accurate enough to be treated as “ground truth” due to the usage of dense MoCap markers. Newly added datasets in our work, *i.e.*, NTU RGB+D and HDE dataset, have their 2D pose estimated from OpenPose [11].

the same pair of RGB and depth images. To align RGB and depth image pixel-wisely, we further warp the RGB image with the estimated projective transform. To distinguish if the depth from a pixel is from the person or the background, we used a pre-trained Mask-RCNN model [149] to segment the person from the background. The segmentation mask is further used for computing $\delta(i, j)$ in Eq. 6.2. Through this process, we acquired millions of pixel-wise aligned RGB-and-depth-image pairs, which are usually directly available from the sensor API. In addition to the RGB-and-depth-image pair, we will still need 2D joints and an initialization of SMPL parameters to start optimizing the objective in Eq. 6.3. We used an off-the-shelf 2D pose estimation model, OpenPose [11], to generate 2D pose J_{est} . We used the pre-trained model from [12] to predict SMPL parameters θ and β . Specifically, we used the same weight for λ_θ , λ_a , and λ_β following [12]. As for λ_d , we empirically set it to 8,000 so that the scale of the loss is the same as others. In the end, we curated a dataset with 680,932 RGB images containing one human subject in the frame, together with corresponding ground truth SMPL parameters.

In addition to the NTU RGB+D dataset, we also included Human3.6M [123], MPI-INF-3DHP [150], LSP [151], LSP-Extended [152], MPII [136], and COCO [153] as our training datasets following the same setup as [12]. Note that we sample 30% from Human3.6M, 30% from NTU RGB+D, 10% from MPI-INF-3DHP, and 30% from the rest in-the-wild datasets

during each training epoch. The number of samples that are gone through in each training epoch is equal to the number of images in the largest dataset (680, 932 here). We give a brief overview of each dataset below, but we refer readers to the original papers for details. Table 6.1 shows statistics and available ground-truth types of these datasets.

Human3.6M: The dataset is collected in a lab-setting indoor environment with human subjects performing different actions like walking, sitting, and phone calling. 3D human joints ground truth is collected by a marker-based MoCap system. Following the same preprocessing steps in [12] and the typical training/evaluation splitting protocol [97]. SMPL parameters are estimated as ground truth based on 67 MoCap markers using MoSh [10]. Given the dense MoCap markers used in MoSh, it’s the most reliable way to estimate SMPL parameters so far (~ 2 cm error per joint). Note that those SMPL parameters estimated by MoSh are available for training in [12], while in our work they are not due to licensing issues. Therefore, we generated pseudo ground truth of SMPL parameters with the pre-trained model in [12] for the purpose of training.

MPI-INF-3DHP: The dataset is also captured in a lab-setting indoor environment (a green screen studio) with human subjects performing different actions. 3D pose ground truth is collected by a marker-less MoCap system. SMPL parameters are estimated as ground truth based on 3D pose ground truth using SMPLify [127].

LSP, LSP-Extended, MPII, and COCO: These datasets are commonly used for 2D pose estimation. Images from these datasets are in-the-wild and more diverse compared with indoor lab-setting ones.

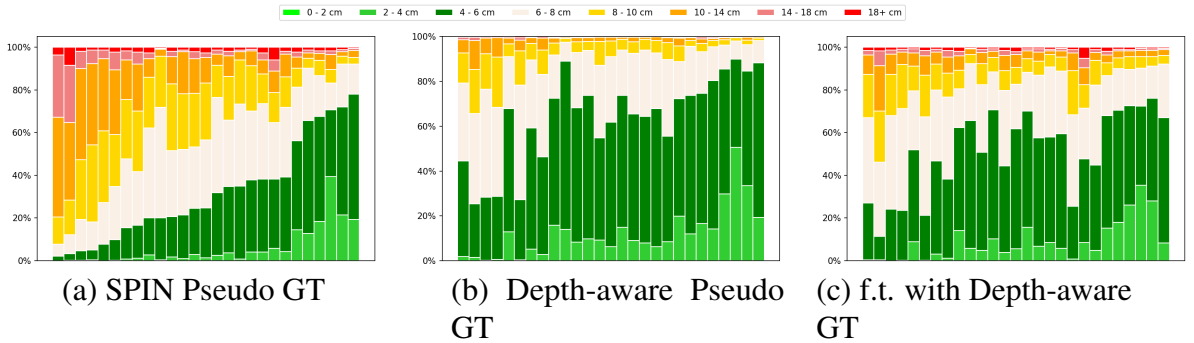


Figure 6.2: Distribution of absolute error of rendered depth under different camera perspectives. Each bar in the figure corresponds to a specific camera perspective. SPIN Pseudo GT is more sensitive to camera perspectives compared with depth-aware GT and the prediction of our fine-tuned model.

Additional data type needed	SMPL oracle	RGB	2D pose oracle	Depth
SPIN GT	✓	✓	✗	✗
SMPLify GT	✓	✓	✓	✗
Depth-aware GT	✓	✓	✓	✓

Table 6.2: Experiment setup comparison. An oracle SMPL parameters initialization and 2D pose can be easily available from any pre-trained models, while RGB and depth images can be directly recorded from various mobile devices with a depth sensor.

6.4.2 Evaluation Datasets and Metrics

Following [12], we evaluate the goodness of fitted SMPL model with common metrics for 3D pose estimation: mean per joint position error (MPJPE) and PA-MPJPE. MPJPE is the mean Euclidean error between each pair of predicted and ground truth 3D joints, while PA-MPJPE is essentially MPJPE after aligning predicted 3D pose with the ground truth 3D pose through Procrustes analysis. Specifically, we report results on Human3.6M [123], MPI-INF-3DHP [150], and 3DPW [143] for this line of evaluation. The evaluation focuses on the accuracy of sparse human joints and cannot reflect the accuracy of human shape and local details of the estimated human mesh.

Another line of evaluation is conducted on foreground-background and six-part segmentation on the LSP dataset [140]. Accuracy and F1 scores are reported for both foreground-background and part segmentation. These metrics are fine-grained indicators of the fitness of the SMPL model across the two dimensions of the image plane. These metrics can indicate the accuracy of human shape on the image plane, but do not capture the local details along the principal axis of the camera.

In our work, we propose another metric to measure the fitness of the SMPL model along the depth direction, which is orthogonal to the image plane. Specifically, we measure the rendered depth of a predicted SMPL model. Following the evaluation of depth estimation, we report the mean absolute error (MAE) and accuracy under different tolerated error thresholds over a human depth estimation (HDE) dataset collected in [154]. In the dataset, human subjects are captured by a Kinect V2 camera while performing different activities in a lab-setting environment. Note that HDE dataset is less challenging than NTU-RGBD dataset because the camera is mounted with a fixed height and distance to the subject and most images are capturing the front view of the subject. Given the dataset is never used for training, we use its training set for the evaluation. Note that the dataset provides the human segmentation mask and aligned RGB and depth image pairs. We adopt the same process as the NTU-RGBD dataset

	Accuracy			MAE (cm)
	1.25 cm	2.5 cm	5.0 cm	
pre-trained [12]	14.0	26.4	46.1	8.21
f.t. with SPIN GT	14.5	27.4	47.6	7.81
f.t. with SMPLify GT	12.8	24.4	43.0	9.14
f.t. with Depth-aware GT	15.8	30.1	52.6	6.67
Depth-aware GT (upper bound)	17.8	33.7	57.7	5.80

Table 6.3: Ablation study: Evaluation on the rendered depth for NTU-RGBD test set.

	Human3.6M		3DPW		MPI-INF-3DHP	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
pre-trained [12]	62.2	41.8	96.9	59.3	103.7	67.9
f.t. with SPIN GT	60.9	41.6	96.9	60.6	102.2	66.5
f.t. with SMPLify GT	60.6	41.5	98.7	62.7	104.2	68.0
f.t. with Depth-aware GT	60.8	41.1	95.4	60.9	100.8	66.2

Table 6.4: Evaluation on the reconstructed 3D pose over different datasets. Both MPJPE and PA-MPJPE has the unit of millimeter.

when generating depth-aware SMPL parameters.

6.4.3 Experiment Setup

We mainly compare the impact of introducing NTU-RGBD dataset in the training process. Specifically, we have generated three different versions of pseudo ground truth for NTU-RGBD dataset, *i.e.*, SPIN GT, SMPLify GT and Depth-aware GT. SPIN GT is directly inferred from the pre-trained model in [12]. Given the estimated 2D pose from OpenPose [11], SMPLify GT is further optimizing SMPL parameters using SMPLify [127]. Given both 2D pose and depth image, Depth-aware GT is further optimizing SMPL parameters using our proposed SMPL parameters optimization approach. The former two versions of ground truth serve as the baselines for our proposed approach. SPIN GT is essentially in the same spirit of NeuralAnnot [141]. We summarize the additional data type needed for generating each version of the pseudo ground truth in Table 6.2. We also remark an oracle SMPL parameters and 2D pose are easy and inexpensive to acquire from a pretrained model. In other words, our approach can be readily adopted on any RGB-D image, which will be ubiquitous as we are witnessing more mobile devices equipped with depth sensor.



Figure 6.3: Qualitative results of SPIN GT and ours Depth-aware GT. The upper panel examples are from HDE dataset and the lower panel ones are from NTU-RGBD test set. For each panel, the first row is the original RGB image. The second and third row are the front view of the human mesh from SPIN GT and Depth-aware GT. The forth and fifth row are the side view of the human mesh from SPIN GT and Depth-aware GT.

	FB Seg.		Part Seg.	
	acc	f1	acc	f1
pre-trained [12]	91.79	0.870	89.36	0.683
f.t. with SPIN GT	92.07	0.875	89.65	0.692
f.t. with SMPLify GT	92.04	0.874	89.66	0.692
f.t. with Depth-aware GT	92.17	0.877	89.76	0.696

Table 6.5: Evaluation on the reprojected pose and shape over LSP test set. Reconstructed human mesh is reprojected to the image plane to evaluate foreground-background and six-part segmentation.

6.4.4 Ablation Study

We evaluate how much a data-driven model can benefit from pseudo ground truth generated with our proposed approach. Since NTU-RGBD dataset is more challenging than HDE dataset, we split NTU-RGBD dataset into training set (84%, 571,537) and test set (16%, 109,395). Our split protocol ensured that samples from the same subject belong to the same set. In other words, we have 80 subjects in the training set and 26 subjects in the test set. We fine-tune the pretrained model from [12] during each training epoch with the sampling strategy mentioned in Section 6.4.1. NTU-RGBD training set is used in the training process and we evaluate the rendered depth on NTU-RGBD test set. Specifically, we report accuracy and MAE performance on rendered depth from the pre-trained model, and fine-tuned models trained with different versions of SMPL pseudo ground truth. As shown in Table 6.3, the data-driven model fine-tuned with our pseudo ground truth achieves the best performance over two other baselines and the pre-trained model itself. By averaging, using Depth-aware GT for finetuning reduces 1.54 cm MAE from the pre-trained model, while MAE itself has the scale of ~ 5 cm. Note that, given the additional depth images, our proposed approach can generate depth-aware ground truth on NTU-RGBD test set with only 5.16 cm MAE, which is essentially the upper bound of a data-driven model can achieve. Recall that NTU-RGBD dataset is collected with various different setup of the camera’s height and distance to the subject. We visualize the distribution of the absolute error of rendered depth on NTU-RGBD test set in Fig. 6.2 under different camera setups. The distributions under different setups vary a lot for SPIN GT. However, they are more consistent for Depth-aware GT, which leverages the depth image and conduct the depth-aware optimization. Similar trend of consistency can also be found in our model fine-tuned on NTU-RGBD training set together with Depth-aware GT. This suggests our fine-tuned model has better generalizability for different camera perspectives.

	Accuracy			MAE
	1.25 cm	2.5 cm	5.0 cm	
pre-trained [12]	18.2	33.8	57.1	5.93
f.t. with SPIN GT	19.1	35.5	59.7	5.62
f.t. with SMPLify GT	18.4	34.1	56.6	6.40
f.t. with Depth-aware GT	19.6	36.2	60.1	5.59
Depth-aware GT (upper bound)	20.3	37.8	62.6	5.16

Table 6.6: Evaluation on the rendered depth over HDE dataset. Our approach outperforms both the pre-trained model [12] and the models fine-tuned under the same semi-supervised setup while different pseudo ground truth.

6.4.5 Quantitative Results

We further leverage all datasets listed in Table 6.1 except HDE dataset to fine-tune the CNN model [97] for human mesh reconstruction. We use the full set of NTU-RGBD dataset for training and its different versions of pseudo ground truth, i.e., SPIN GT, SMPLify GT and Depth-aware GT. We denote the model that is fine-tuned from the pre-trained model as “f.t. with SPIN GT”. Similar notation applies to other versions of pseudo ground truth. We report evaluation results on various human pose related computer vision tasks, *i.e.*, 3D pose estimation, foreground-background and six-part segmentation, and human depth estimation, in Table 6.4, 6.5, 6.6 respectively. Models that fine-tuned on our pseudo ground truth are able to push the performance on various human pose related tasks to the new limit compared with the state-of-the-art pre-trained model [12].

6.4.6 Qualitative Results

To qualitatively demonstrate the benefit and impact that is brought by depth, we visualize and compare human meshes between those that are predicted from the pre-trained model and those that are optimized by our proposed approach. Figure 6.3 shows several examples in HDE dataset and NTU-RGBD test set. Both meshes have a consistent front view with the corresponding RGB images, while baseline meshes seem to be overly thin. Our meshes are much more visually reasonable than the baseline meshes not only for cases that have significant erroneous error along the depth dimension like the forth column in the upper panel, but also for cases that have limited amount of error like the seventh column in the upper panel. We highlight a good human mesh like the last column in the upper panel could be important for downstream human pose related applications like human bodily expression recognition [101]

because relaxation is “encoded” in our reconstructed human mesh.

6.5 Summary

This chapter presented an optimization-based approach to absorb depth information from available RGB-D training data into human mesh reconstruction. We further leveraged the refined human mesh to train a neural network in a semi-supervised fashion. By introducing those refined human mesh in the training process, our experiment has shown the trained network is more depth-aware, robust and accurate, , *i.e.*, outperforming not only along the depth dimension, but also on the image plane. The NTU-RGBD dataset together with pseudo ground truth generated from our proposed approach could be an asset for future human mesh reconstruction research. Because the NTU-RGBD dataset is still collected in a lab setting environment, future work could consider leveraging the approach for large-scale cost-effective RGB-D data collection efforts.

Chapter 7 |

Conclusions and Future Work

7.1 Summary

We proposed a scalable and reliable video-data collection pipeline and collected a large-scale bodily expression dataset, the BoLD. We have validated our data collection via statistical analysis. To our knowledge, our effort is the first quantitative investigation of human performance on emotional expression recognition with thousands of people, tens of thousands of clips, and thousands of characters. Importantly, we found significant predictive features regarding the computability of bodily emotion, *i.e.*, hand acceleration for emotional expressions along the dimension of arousal. Moreover, for the first time, our deep model demonstrates decent generalizability for bodily expression recognition in the wild. We also proposed an optimization-based approach to incorporate depth information into human mesh. We have evaluated the effectiveness of leveraging these improved human mesh to build predictive model for human mesh reconstruction in a semi-supervised learning fashion.

7.2 Future Work

Bodily expression recognition can be approached in a pyramid structure as shown in Figure 7.1. The base layer would be data collected from sensors, *e.g.*, videos from cameras. The third layer is human pose movement extracted from raw data. Depending on the type of the raw sensors, human pose movements are detected with automated algorithms or models. The second layer provides a semantic categorization for human movements and the top layer is bodily expressed emotions. To date, there is still little work on the second layer. An important and promising future direction would be bridging bodily expressed emotions and human pose movements with movement coding. Unlike other layers, the representation of movement coding is not

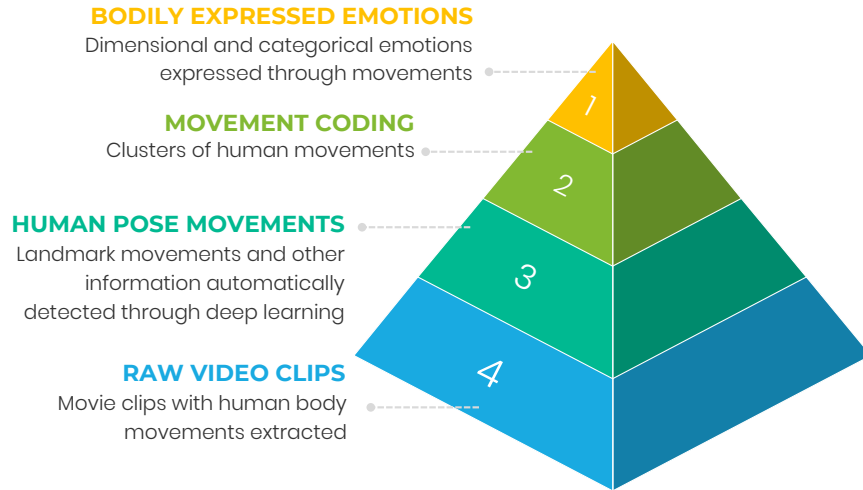


Figure 7.1: Layers of abstraction for bodily expressed emotions.

directly or easily available. It takes collaborations between computer scientists, psychologists and other domain experts like certified movement analyst (CMA) to define a useful movement coding representation. The bridge between layers two and three is more computational because computational models or algorithms should be proposed to automatically extract movement coding from raw videos (layer four) or human poses (layer three). The bridge between layer three and layer four, however, needs more effort from psychology. Our work has shown the effectiveness of Laban Movement Analysis and Laban notation is consequently a good starting point for such a movement coding layer.

Other possible directions for future work are numerous. First, our model’s regression performance of arousal is clearly better than that of valence, yet our analysis shows humans are better at recognizing valence. The inadequacy in feature extraction and modeling, especially for valence, suggests the need for additional investigation. Second, our analysis has identified demographic factors in emotion perception between different ethnic groups. Our current model has largely ignored these potentially useful factors. Considering characters’ demographics in the inference of bodily expression can be a fascinating research direction. Third, our “learning-from-depth” framework for human mesh reconstruction has shown superior effectiveness under various camera perspectives, thanks to the diversified camera perspectives in the original RGB-D dataset. Future work could consider leveraging the approach for large-scale cost-effective in-the-wild RGB-D data collection efforts. Finally, although this work has focused on bodily expression, the BoLD dataset we have collected has several other modalities useful for emotion recognition, including audio and visual context. An integrated approach to study these will likely lead to exciting real-world applications.

Appendix |

IVAR: Image Variation Retrieval via Feature Learning

Detecting image manipulation and near-duplicates has long been an interesting research topic due to applications in image retrieval and copyright protection. Recent advances in deep learning (*e.g.*, photo-realistic image generation, neural style transfer, and automatic masking) provide increasingly more options to dramatically transform image appearance, making it challenging for conventional approaches to retrieve manipulated image variations. In this chapter, we introduce this emerging problem, image variation retrieval, *i.e.*, **IVAR** and comprehensively analyze its challenges using the off-the-shelf CNN features. We treat an image and its manipulated variations as one instance and formulate a feature learning approach that aims to distinguish one instance from others. The central idea is to design a loss function that minimizes intra-instance distances and maximizes inter-instance distances and solve the objective practically using a stratified sampling mechanism. Experimental results show that our approach can effectively retrieve image variations from both a large image database and significantly outperforms off-the-shelf deep learning features.



Figure A.1: Examples of two common image manipulation processes – compositing and retouching. The right three images are manipulated from the first image on the left. Specifically, the second and the third image are edited via compositing, *i.e.*, inserting the man and the woman into the image. The fourth and the fifth image are retouched with different artistic filters.

A.1 Introduction

Recent advances in deep learning (such as photo-realistic image generation, neural style transfer, and automatic masking) dramatically simplify and automate image editing and transformation. As a result, more and more artistically transformed images have been included in all sorts of image databases and personal albums. A common copyright protection challenge is to detect such manipulated images from a large image collection (*e.g.*, images on sale on photography e-commerce sites). Figure A.1 shows some examples of manipulated images. Given an arbitrary one among these images as a query, the problem we investigate is how to efficiently detect other images in a large image collection that are generated from a same source image.

Image variation retrieval is an emerging research problem due to the rapid development of deep learning in image understanding and generation. Potential applications include image hosting services and real-time, mobile apps. For instance, image variation retrieval provides an alternative and straightforward way to organize image collections, either on mobile albums, desktop disks, or cloud collections. With its support, users can easily find different variations of the same image. On the other hand, as gathering variations of the same image gets easier, protecting the copyright of original photographic creations is made easier. The photographer of an original photograph can use the photo to find unauthorized manipulations of the photo on the Internet.

An intuitive solution to the problem is to apply content-based image retrieval techniques, formally instance-level image retrieval, which has long been focusing on retrieving the exact same instance as defined in benchmark datasets, such as objects [155], scenes [156] and semantic categories [157]. Similar images under this conventional setting usually include images of the same object or scene viewed under different imaging conditions. Although off-the-shelf CNN features have showed incredible results under the conventional image retrieval setup, this solution does not apply well to the proposed problem. A primary reason is that manipulated image variations can possibly have a dramatically different visual look (as shown in Figure A.1) compared with the original as they are commonly generated through compositing (*i.e.*, combining image layers of different sources), retouching (*e.g.*, ranges from color or tonal adjustment [158] to neural style transfer [159, 160]) or both. Whereas off-the-shelf CNN features are good at searching for similar objects or scenes, they are not aiming for identifying image variations that have dramatically different looks.

To find an appropriate feature space where variations of the same image can be projected tightly, we propose a feature learning system that aims to learn from a training dataset with the goal of minimizing distances across variations of the same image and meanwhile maximizing

distances across different images in a learned projection space. We are motivated by a metric learning schema, where a loss function is defined to characterize the similarity over sample pairs or triplets. Specifically, in deep learning domain, siamese neural network or triplet network are widely used for various tasks [161–163]. However, such approaches usually require representative similar/dissimilar sample pairs for representation learning. They cannot be efficiently applied to our task because there are so many types of image manipulations especially when considering the combinations of compositing and retouching. Moreover, we are lack of large image database to solve the learning problem. To learn the feature space efficiently, we first propose and build an automated image manipulation pipeline to simulate the process of compositing and retouching. We then formulate the feature learning algorithm given the aforementioned intuition. To practically solve the objectives, we adopt the stratified sampling strategy to densely model every sample pair. We evaluate the proposed approach both quantitatively and qualitatively using both the synthetic and real-world datasets.

Our **main contributions** are as follows:

- We introduce and formulate the problem of image variation retrieval in a large image collection. We simulate various image transformation outputs and create an image database of image variations. By utilizing off-the-shelf CNN features, we comprehensively examine the impact of having transformed images to the representation space.
- We propose a feature learning approach to tackle image variations retrieval problem. In particular, we propose a hinge loss based loss function to model pair-wise representation distance in the embedding space for better representation learning. We adopt the stratified sampling strategy to allow the loss to be efficiently approximated within the mini-batch.
- We adopt various techniques to automatically generate image variants for training. We collected one synthetic dataset and two real-world evaluation datasets to validate the effectiveness of our proposed approach. Our models show superior retrieval performance on all evaluation datasets compared with off-the-shelf CNN features.

A.2 Related Work

The problem of image variation retrieval is related to three research problems: instance-level image retrieval, near-duplicate image detection, and image manipulation detection. We discuss related work in each of them below.

Image Retrieval and Near-Duplicate Image Detection. A number of methods have been developed since the introduction of invariant local features [164–168]. Studies on this topic

commonly follow the pipeline of extracting SIFT features from images and indexing those features into visual words with the Bag-of-Word (BoW) model. Off-the-shelf CNN features are known to be compact and invariant to various transformations and have been widely used in the retrieval area [169]. Our proposed problem of image variation retrieval, however, is fundamentally different from conventional instance-level image retrieval in that we define the instance as different variations of the same image rather than the conventionally formulated one by objects or scenes under a different imaging condition. Meanwhile, the proposed task is beyond the scope of near-duplicate image detection as the latter intends to find different versions of the same image that can be visually identified as the same image. In our framework, image variations can expand with dramatically different appearances.

Image Manipulation Detection. Image manipulation detection aims to localize the tampered part of an image, especially targets detecting the fake regions, *e.g.*, an object is inserted or removed. Conventionally, local noise is used as a cue to distinguish whether an region is original or tampered [170]. Recently effort on image harmonization [171] and image generation [172, 173] makes the cue of local noise less powerful to make a judgment. Studies then focus on jointly leverage the idea of local noise and deep learning to detect unreal image regions [174]. Whereas studies under this topic primarily focus on tampered region localization given a single image, our work address the problem of image variation retrieval, *i.e.*, given one example as a query, we retrieve its variations (including the original copy) in a large image collection.

Our work is also related to studies that investigate robustness of off-the-shelf CNN features on image filters [175] as we both study image distribution on the projected space. Studies in the same line also attempted to take image transformation (*e.g.*, image perturbation [176], adversarial examples [177], and other aggressive data augmentation [178]) as an data augmentation mechanism for classification accuracy boost. Unlike these work that mostly focus on producing a robust classification model by training with a larger volume of data, our work focus on minimizing the distance among image variations in the projected feature space to address image variation retrieval.

A.3 Analysis of Image Transformations

To study the problem of image variation retrieval, we need a large collection of images, where for each image we need several variations that are generated through image editing. Collecting such a real-world dataset is challenging because photo variations are often privately held and not posted in the public. Although some photographers may share publicly one or a

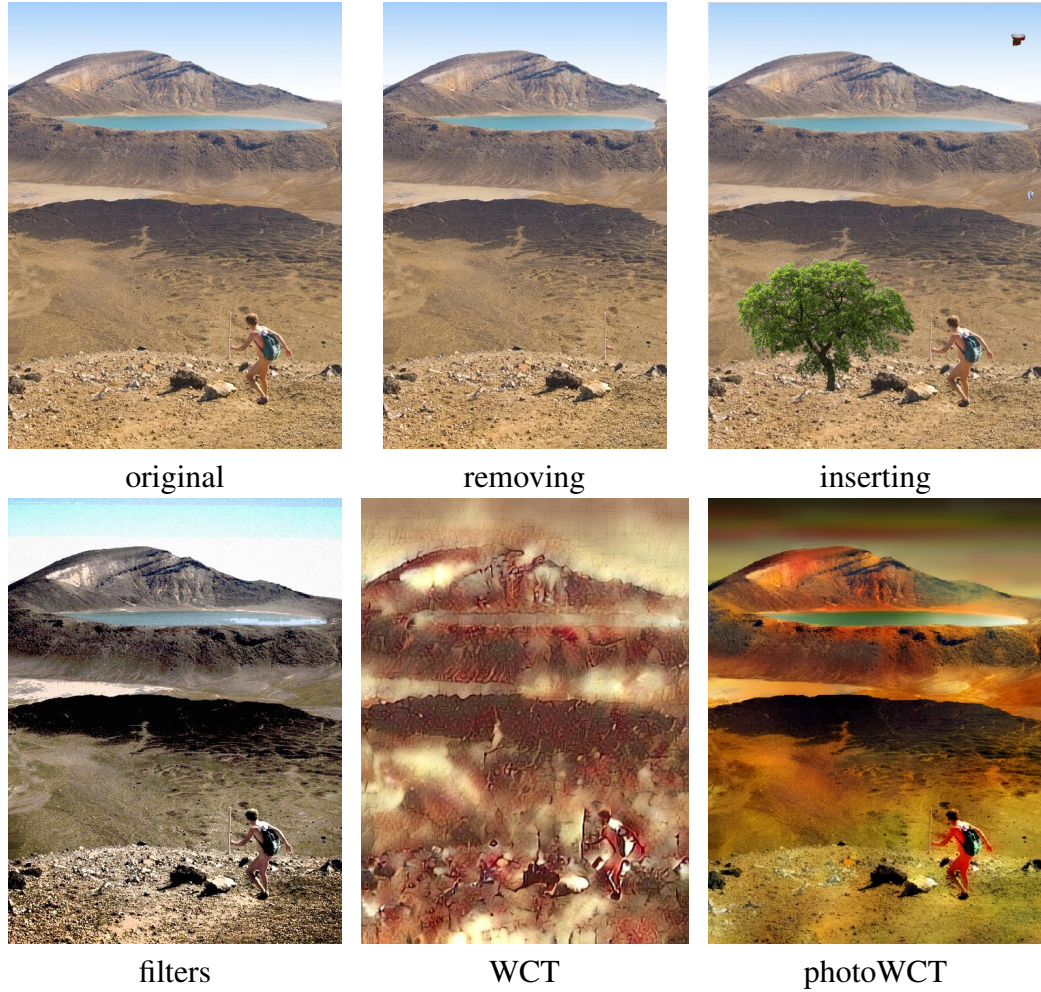


Figure A.2: An instance in our training set. The removing example is generated by removing the boy out of the image. The inserting example is generated by inserting the car. The filters example is generated with a Gaussian blur filter. The WCT and photoWCT examples are generated by two different style transfer algorithms.

few edited images, they often do not make the original unedited images public. To have a deeper understanding of the challenges of the proposed problem, we first create a simulated dataset with representative image transformations, including removing, inserting, filters, neural style transfer (later referred to as WCT), and photo-realistic style transfer (later referred to as photoWCT), as shown in Figure A.2. We then conduct comprehensive analysis on the off-the-shelf CNN feature space. This section first introduces the image manipulation details in Section A.3.1 and present the analysis results in Section A.3.2.

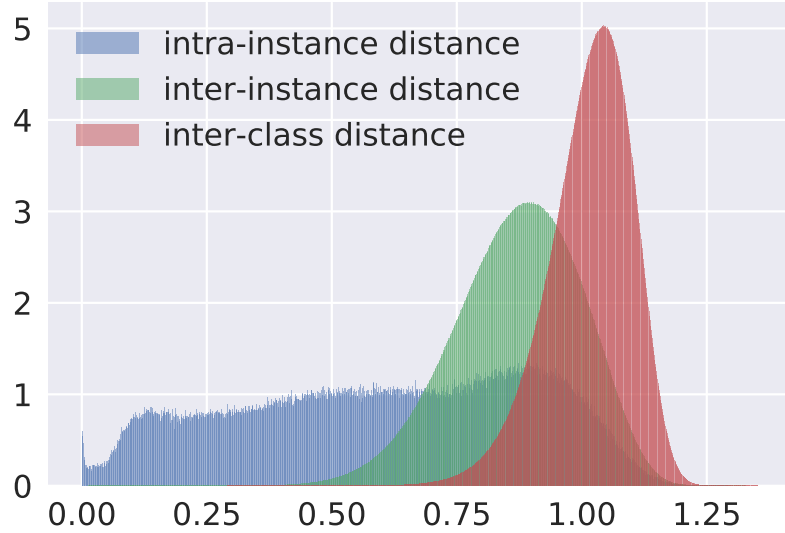


Figure A.3: Distance distribution of baseline model. Histograms of intra-instance distance, inter-instance distance, and inter-class distance on the off-the-shelf CNN feature space.

A.3.1 Simulated Image Transformations

As we are working on retrieve image variations, scene images are more appropriate compared with images dominated by objects. We thus choose the commonly used Places dataset [179] as an original image collection. To mimic compositing, we pick the two most basic operations, *i.e.*, inserting and removing objects. Specifically, we insert objects taken from MS-COCO dataset [153] based on the provided segmentation masks. To avoid inserted images dominating the image, inserted objects are scaled so that the total area of inserted objects is between $\frac{1}{4}$ and $\frac{1}{3}$ of the area of original image. The number of inserted objects in an image ranges from one to five. To remove objects from the original image, we first generate object masks by running mask-rcnn [149]. We then randomly select one object to remove via seam carving [180]. Examples of inserting and removing can be found in Figure A.2. On retouching, we automate three types of edits, *i.e.*, filters, neural style transfer, and photo-realistic style transfer. Filters include Vintage, Lomo, Clarity, Sin City, Sunrise, CrossProcess, Orange Peel, Love, Grungy, Jarques, Pinhole, Old Boot, Glowing Sun, Haze Days, Her Majesty, Nostalgia, Hemingway, Concentrate¹ and random combinations of basic image processing blocks, such as, hue, saturation, exposure adjustment, unsharp, denoising and so on. While there are many style tranfer related work, we pick two recent ones, WCT [159] and photoWCT [160]. The first one generates images with painterly effects and the latter ones generate photo-realistic results. For WCT, style images are

¹<http://jason2506.github.io/imeffect/>

randomly picked from a large painting dataset² (100k+ paintings). For PhotoWCT, we use the same set of painting images as style images.

Following the above process, we create Places-IVAR-Train and Places-IVAR-Test dataset with images from training and validation set of Places dataset. The original training and validation set contains about 180 million images and 36,500 images respectively. Each original image has five manipulated variants in our datasets. We further use Places-IVAR-Test in our analysis as the number of images is more manageable while Places-IVAR-Train is used as training data for our approach.

A.3.2 Analysis of Image Transformations

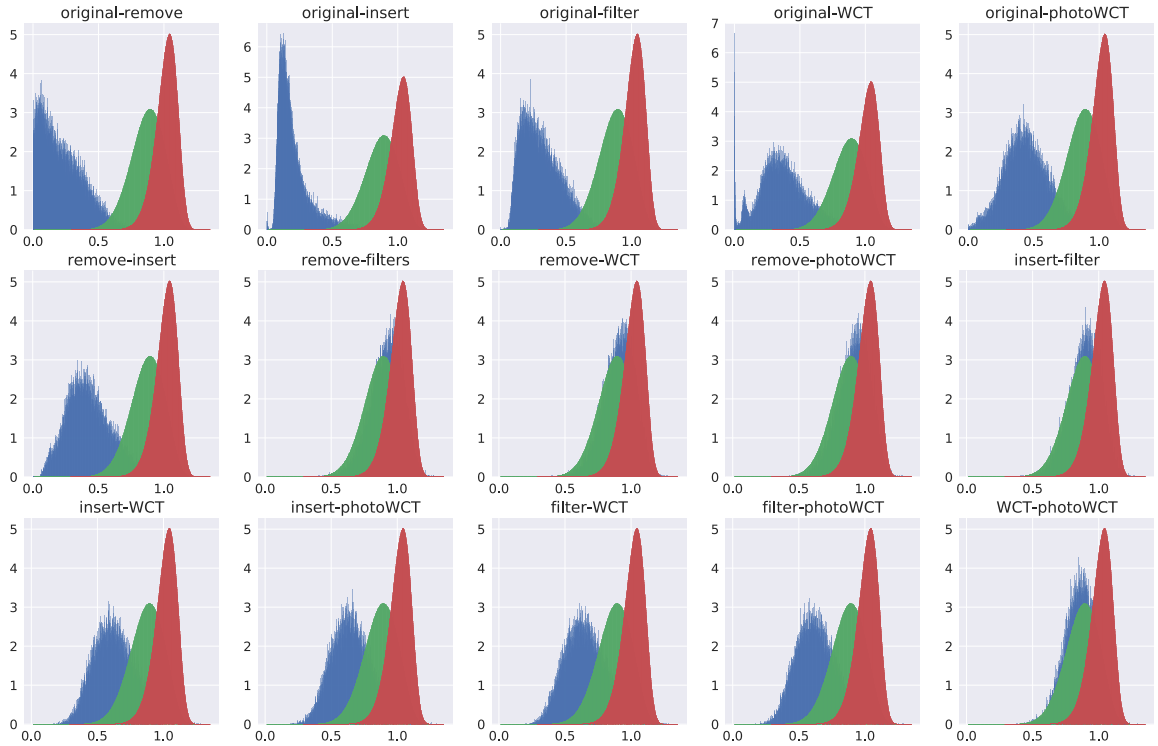


Figure A.4: Intra-instance distance distribution over different transformations. The blue histogram is the distribution of intra-instance distance between pairs in the title, *e.g.*, raw and remove for the first figure. The green histogram is the distribution of inter-instance distance. The red histogram is the distribution of inter-class distance. The farther intra-instance distance is from the other two, the more possible that a model retrieves desired image variants.

To understand the image distribution changes brought by introducing transformed images,

²<https://www.kaggle.com/c/painter-by-numbers>

we statistically analyze image distributions on the off-the-shelf CNN feature space. We denote by instance an original image and all its variations and denote the three type of distances as following: the distance distribution among image of the same instances (*i.e.*, intra-instance distance), images of different instances (*i.e.*, inter-instance distance), and images of different semantic categories (*i.e.*, inter-class distance). We then statistically analyze the three distances on the Places-IVAR-Test dataset. This study explicitly indicates that the intra-instance variance is far larger than the inter-instance variance and inter-class variance. Meanwhile, the distribution of intra-instance distance largely overlap that of inter-instance distance. This concludes the reason why off-the-shelf CNN features do not effectively distinguish different instances.

As shown in Figure A.3a, we plot the distribution of intra-instance, inter-instance and inter-class distance on Places-IVAR-Test dataset with our baseline model. The baseline model is a ResNet50 model pretrained on Places dataset for scene classification. We extracted and normalized the activation before fully connected layer as image representation (2048-D feature vector). Note that euclidean distance between any two feature vectors ranges from 0 to $\sqrt{2}$ ³. Within each instance, we compute euclidean distance over all pairs of image variations, *i.e.*, $\binom{6}{2} = 15$, and plot the histogram of intra-instance distance. For different instances, we also enumerate all pairs of images and separate them into inter-instance distance and inter-class distance based on the consistency of their class label. Due to the large number of inter-class pairs, we only samples 166 million pairs.

To explore the impact of individual edit on the distance distribution, we plot decomposed intra-instance distance into specific variant type pairs. Specifically, we collect and decompose above computed intra-instance distance over each variant pair. Note that for variant pair like remove-filters, one can deem the filter variant is generated from the remove variation with two processes, *i.e.*, insert and filter. As shown in Figure A.4, we find that original images and other variants is still separable on the off-the-shelf CNN feature, *e.g.*, original-remove, original-filter. However, distance between images that are transformed across compositing and retouching, *e.g.*, remove-photoWCT, insert-filter, largely overlap with inter-instance and inter-class distance. This suggests retrieving an image transformed by a combination of compositing and retouching from original image is more challenging for the pretrained model, compared with retrieving a composited or retouched image. These results also indicate that retrieval among manipulated images is challenging on the off-the-shelf feature space.

³ $d = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\mathbf{x}^2 + \mathbf{y}^2 - 2\mathbf{x}\mathbf{y}} = \sqrt{2 - 2\mathbf{x}\mathbf{y}}$, $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$. Our feature vector follows a ReLU layer. Hence, $x_i > 0, y_i > 0, \forall i$. and $\mathbf{x}\mathbf{y} \in [0, 1]$.

A.4 The Approach

In this section, we present the proposed image variation retrieval system. Given an image set $M = \{M_i\}$ and a query q , our goal is to retrieve for q 's variations $\{q_v\}$ in M , where cardinality of $\{q_v\}$ is unknown. We follow the convention of image retrieval and assume there's a feature space f , we compute the projection of both q and $m \in M$ on the feature space, i.e., $f(M)$ and $f(q)$. We denote by $d_i(f(m_i), f(q))$ pair-wise euclidean distance between m_i and q . Following [181], we use $L2$ normalized feature vector to compute the distance⁴. We rank $m \in M$ according to $\{d_i\}$ and derive the retrieval results.

To resolve the challenges experienced by the off-the-shelf CNN feature space as we presented in Section A.3.2, the key question is how to get the feature space f so that the query q and its variations $\{q_v\}$ could stand closer on the space f . Instead of handcrafting a feature space, we propose to learn this feature space based on a training dataset. We denote by $X = \{x_i^u\}$ the image set for training where x_i^u refers to an image that transformed from i -th sample via transformation u . We denote by N the total number of samples and denote by T the total number of transformations, so the cardinality of X is $N \times T$.

To learn a feature space, the central idea is to minimize the intra-instance distance and maximize inter-instance distance. As presented in Figure A.5, we denote by d_+ the intra-instance distance, i.e., the distance between x_i^u and x_i^v and denote by d_- the inter-instance distance, i.e., the distance between x_i^u and x_j^v , where $i \neq j$. We define this function as following:

$$\begin{aligned} \mathcal{L}_u = & \frac{1}{T^2 \binom{N}{2}} \sum_{\substack{\forall u, v \in S; \\ i < j, \forall i, j \in I;}} \max(0, \beta_1 - d_+) \\ & + \frac{1}{N \binom{T}{2}} \sum_{\substack{\forall i \in I; \\ u < v \forall u, v \in S}} \max(0, d_- - \beta_2) \end{aligned} \quad (\text{A.1})$$

where $f(x; \theta)$ is a parametric model with parameters θ that project x onto a feature space, β_1 and β_2 are the margins, and ideally, we hope d_+ is smaller than β_1 and d_- is larger than β_2 . Distance in the equation is measured by euclidean distance:

$$\begin{aligned} d_+ &= \|f(x_i^u; \theta) - f(x_i^v; \theta)\|_2^2, \\ d_- &= \|f(x_i^u; \theta) - f(x_j^v; \theta)\|_2^2, \end{aligned} \quad (\text{A.2})$$

Note that we compute $\mathcal{L}_u$ by enumerating all pairs among $N \times T$ images, i.e., $\binom{N \times T}{2} =$

⁴Assume unnormalized feature vector is $\hat{f}$, then $f = \hat{f} / \|\hat{f}\|_2$

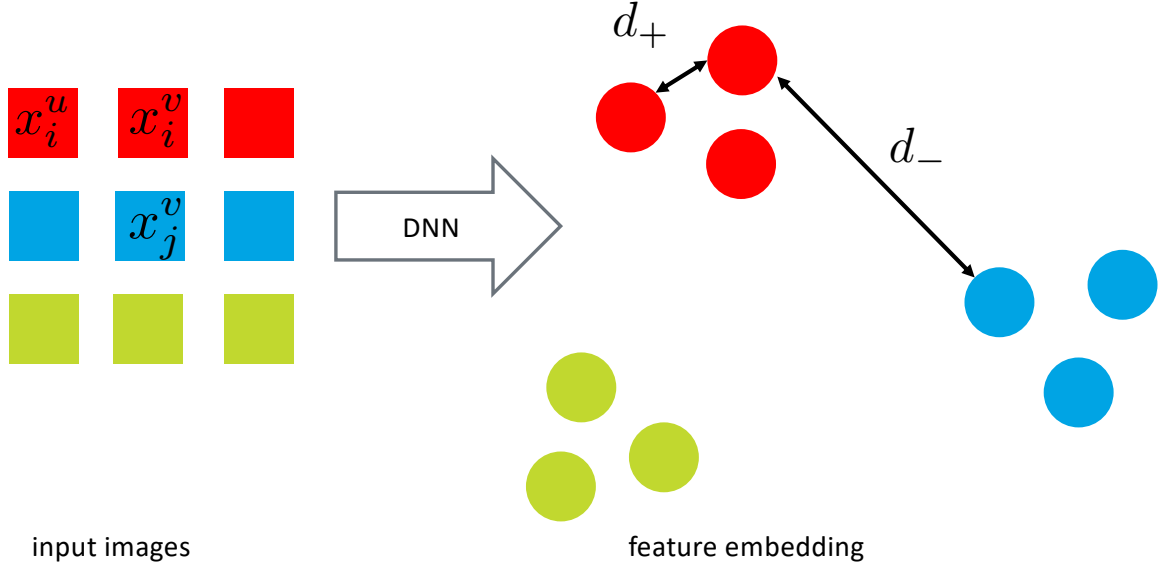


Figure A.5: Objective illustration. Squares with the same color are from the same instance. Our goal is to increase inter-instance distance d_- and decrease intra-instance distance d_+ .

$$T^2 \binom{N}{2} + N \binom{T}{2}.$$

Since semantic categories are still an important aspect of our task, we also incorporate standard cross entropy loss in the original classification task.

$$\mathcal{L}_{ce} = \frac{\sum_{i,u} \gamma_i^u \sum_c \mathbb{I}(y_i = c) \log(p_i^u(c))}{\sum_{i,u} \gamma_i^u + \epsilon} \quad (\text{A.3})$$

where $y_{i,u}$ is the class prediction of x_i^5 , $p_i^u(c)$ is the probability of example x_i^u being predicted with class c , γ_i^u controls the weight of cross entropy of sample x_i^u , ϵ is a small number to prevent denominator to be zero, and $\mathbb{I}(x) = 1$ iff x is true.

Considering both $\mathcal{L}_u$ and $\mathcal{L}_{ce}$, our objective is:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \mathcal{L} \quad (\text{A.4})$$

where

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \mathcal{L}_u \quad (\text{A.5})$$

We use α to adjust the contribution of the two losses.

We use mini-batch stochastic Gradient Descent (SGD) to solve Equation A.4. However, enumerating all pairs of images in the dataset at the same time is computationally expensive and impractical due to limited GPU memory. To solve Equation A.4 practically, We approximate it

⁵ $y_{i,u} = y_i$

Datasets	Places-IVAR-Test	CASIA	MIT-Adobe5k
# instance	36,500	1,955	5,000
# query image	219,000	7,997	30,000
# additional distractor images	0	495,719	433,734
# total number of images	219,000	503,716	463,734
compositing by inserting	✓	✓	
compositing by removing	✓		
retouching with filters	✓		✓
retouching with style transfer	✓		

Table A.1: Evaluation dataset statistics.

by enumerating all pairs of images within mini-batch. As SGD requires that each sample in the mini-batch is uniformly sampled from the dataset, in this work, we draw pair-wise samples through stratified sampling.

For a mini batch of size $n \times t$, we start with uniformly sampling n samples and then uniformly sample t variations of each sample. Comparing with uniform sampling across the $N \times T$ examples in X , by applying the stratified sampling strategy, we are able to sample sufficient image pairs of the same instance in each mini-batch. The usage of stratified sampling significantly boost the training efficiency and make Equation A.4 practically solvable.

A.5 Experiments

We evaluate the proposed approach **IVAR** on both synthetic dataset and real world datasets by examining both numeric metric and visual results. This section introduces the dataset we utilized, the baselines that we compared with, and implementation details, our experimental settings and results.

A.5.1 Datasets

As scene images are more appropriate for the proposed image variation retrieval task, we manipulated images as presented in Section A.3.1 created Places-IVAR-Train and Places-IVAR-Test using the Places dataset [179]. It results in 18 million instances in Places-IVAR-Train. As one instance includes 6 variations, Places-IVAR-Train contains 108 million images in total. Meanwhile, Places-IVAR-Test includes 36,500 instances and 219K images. We use Places-IVAR-Train to learn feature space and evaluate the performance of image variation retrieval on Places-IVAR-Test. In Places-IVAR-Test, every image can be used as a query, so it ends up with

Model	Settings	Places-IVAR-TEST	CASIA	MIT-Adobe5k
Baseline		70.77 (55.088)	92.01	99.57
Ours- β ($\alpha = 50$)	$\beta_1 = 0.3, \beta_2 = 0.6$	97.14 (55.866)	91.76	99.84
	$\beta_1 = \mathbf{0.4}, \beta_2 = \mathbf{0.6}$	97.87 (55.858)	92.17	99.85
	$\beta_1 = 0.4, \beta_2 = 0.7$	97.20 (55.989)	92.11	99.85
	$\beta_1 = 0.3, \beta_2 = 0.7$	97.28 (55.871)	91.74	99.84
	$\beta_1 = 0.4, \beta_2 = 0.8$	97.40 (55.915)	91.82	99.84
	$\beta_1 = 0.3, \beta_2 = 0.8$	97.23 (55.707)	91.59	99.81
Ours- α ($\beta_1 = 0.4, \beta_2 = 0.6$)	$\alpha = 0$	94.94 (56.044)	91.91	99.82
	$\alpha = 10$	96.71 (55.959)	92.46	99.84
	$\alpha = 50$	97.87 (55.858)	92.17	99.85
	$\alpha = 100$	98.25 (55.803)	91.85	99.84

Table A.2: Results on evaluation dataset. Image variation retrieval results in terms of mAP are reported on three evaluation datasets. On Places-IVAR-Test, top 1 classification accuracy is also reported in the bracket.

219K query images.

In order to evaluate **IVAR** in a real-world settings, we set up two additional datasets, *i.e.*, **CAISA** and **MIT-Adobe-5k**. **CAISA** is used by [182] for tampered image detection and **MIT-Adobe-5k** [158] is created for personalized automatic tonal adjustment. In **CAISA**, the tampered image are generated by manually inserting objects and one instance on average has 4 variations, which results in 1, 995 instance and 7, 997 query images. In **MIT-Adobe-5k**, image variations are created by professional photographers through retouching (e.g., filters). In total, **MIT-Adobe-5k** has 5K instance and 50K query images.

As these two datasets contain relatively small number of images, to evaluate the retrieval performance over large scale image database, we collect additional distractor images by querying all images in each of the three datasets with Google Image⁶ and downloads top related images. This approach significantly increase the total number of images and we summarize statistics of the four evaluation datasets in Table A.1.

A.5.2 Evaluation Metric

We use standard mean average precision (mAP) to evaluate retrieval performance [168]. Since a different instance could have different number of variations, we take the average of average

⁶<https://images.google.com/>

precision over instance instead of query images. Specifically, we compute mAP by:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N \frac{1}{t_i} \sum_{j=1}^{t_i} \text{AP}_{ij} \quad (\text{A.6})$$

where N is total number of instances, t_i is the number of variations of the i -th instance and AP_{ij} is average precision of j -th image variant in i -th instance. Since semantics are also an important aspect for image variant retrieval, we also report classification accuracy (top-1) on Places-IVAR-TEST. To compare with the off-the-shelf CNN model, we only evaluate the classification accuracy on original images in the Places-IVAR-Test (i.e., 36, 500 images).

A.5.3 Implementation Details

We use our Places-IVAR-Train to train our model. Since each instance has 6 images, we set t as 6 during stratified sampling. Our model is fine-tuned over pretrained scene classification model [179]. Our model architecture is ResNet50, and we use 80 instances for a batch, i.e., total batch size is 480. We use the activation before fully connected layer as image representation (2048-D feature vector). All images are augmented with scaling, flipping and affine transformation during training. The learning rate starts from 0.001 and is divided by 2 every 2 epochs. Our experiments are conducted on four Nvidia Tesla P100 machine. Training is done after 20 epochs.

A.5.4 Experimental Results

As we are using ResNet50 network architecture, to make a fair comparison, we use the off-the-shelf ResNet50 model pretrained on Places dataset [179] as the baseline. We compared our approaches with the baseline on all the three evaluation datasets on mAP.

Table A.2 summarizes the results of the proposed approaches under different settings. As shown in the Table, the proposed approach outperforms the baseline over all evaluation datasets. In particular, our approach has significantly improved the retrieval performance on Places-IVAR-Test by more than 25%. On the CASIA and MIT-Adobe5k datasets, while the baseline shows very high mAP, our fine-tuned models further reduce about 65% error in MIT-adobe5k and 6% error in CASIA compared with the baseline⁷.

The primary reason why the proposed approach did not improve much on the CASIA and MIT-Adobe5k dataset is because in both of the datasets only one type of image manipulation was used respectively. In CASIA, only inserting was used and in MIT-Adobe5k, only filter was

⁷For CASIA, $1 - \frac{100-92.46}{100-92.01} = 6\%$; for MIT-Adobe5k, $1 - \frac{100-99.85}{100-99.57} = 65\%$.

applied. According to our analysis presented in Section A.3.2, the off-the-shelf CNN feature space separates reasonably well the intra-instance distance vs. inter-instance distance. Our experimental results align with our aforementioned analysis conclusion.

To determine the optimal β_1, β_2 and α , we perform two sets of ablation studies. In **Ours- β** , we let $\alpha = 50$ to compare different values of β_1 and β_2 . In **Ours- α** , we use the optimal β_1 and β_2 out of the **Ours- β** setting and examine the effect of the proposed hinge loss on the retrieval performance. In both **Ours- β** and **Ours- α** , we have $\gamma_i^u = 1$.

As shown in Table A.2, in Ours- β , our fine-tuned model shows similar mAP under different β_1 and β_2 . This indicates that the proposed approach is not sensitive to β . In Ours- α , $\alpha = 100$ results in the highest mAP value on Places-IVAR-TEST, $\alpha = 10$ leads to the highest mAP value on CASIA and $\alpha = 50$ results in the highest mAP value on MIT-Adobe5k. As they are all higher than the feature space trained with $\alpha = 0$, the results across three evaluation datasets consistently show the effectiveness of the hinge loss $\mathcal{L}_u$.

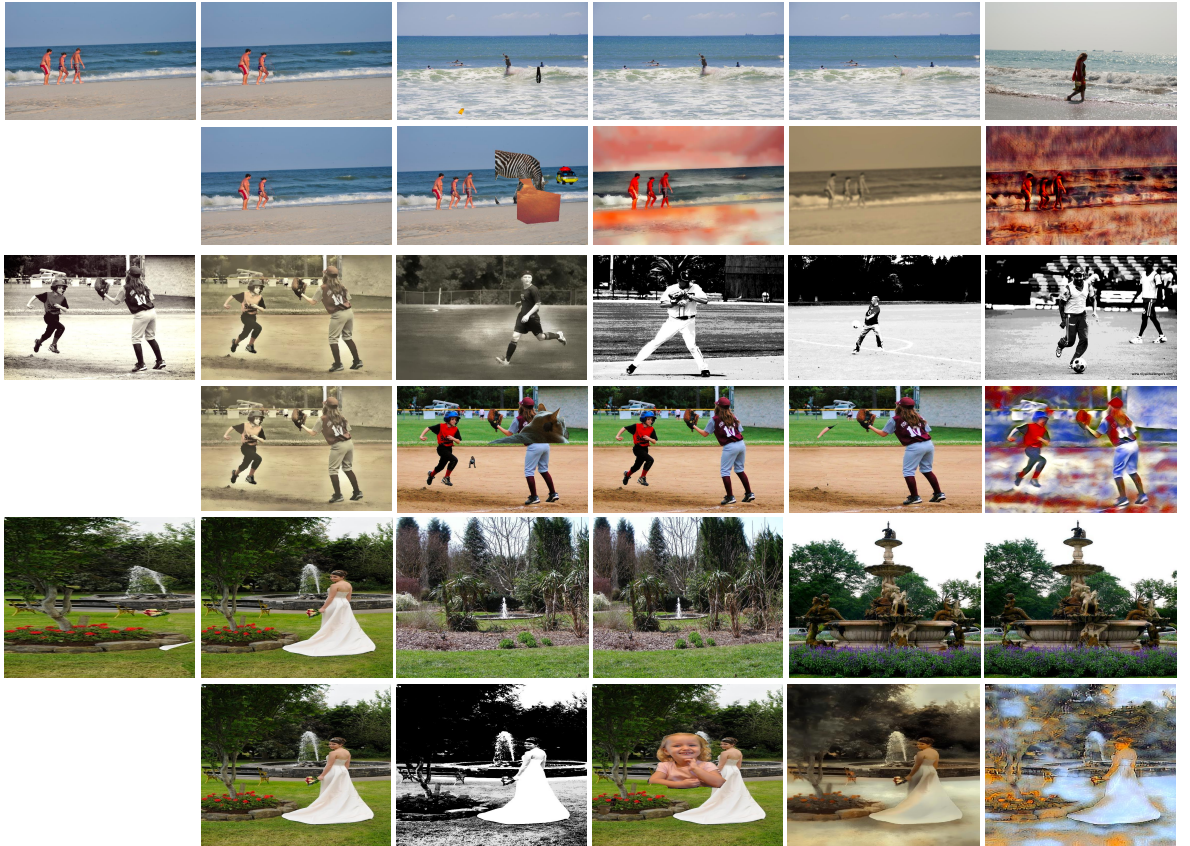


Figure A.6: Example retrieval results on the Places-IVAR-Test dataset. The first column is the query image. For each two rows, the first row is retrieved using the baseline model and the second row is retrieved using our model. Similarity decreases from the left to the right.

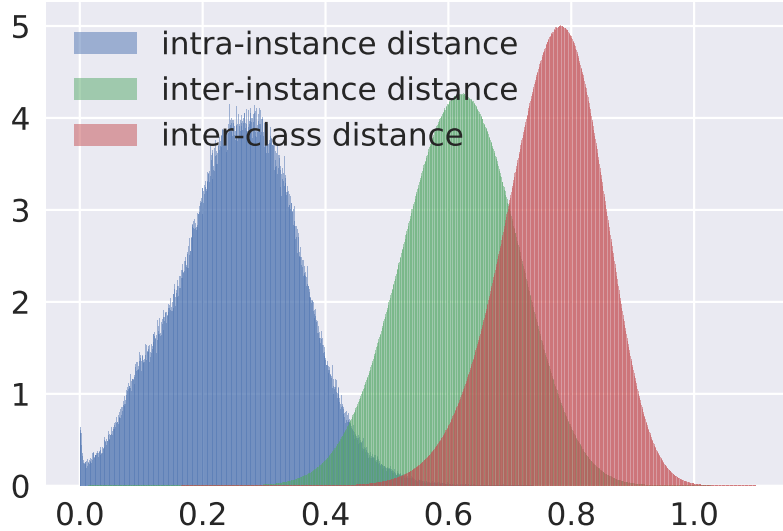


Figure A.7: Distance distribution of our model. Histograms of intra-instance distance, inter-instance distance, and inter-class distance on learned feature space.

Qualitatively, we visualize retrieval results using the baseline and using the proposed approach on Places-IVAR-TEST and results in Figure A.6. The images on the left are the query images, the first row presents the baseline results and the second row presents the results with the proposed approach⁸. As shown in the figure, our approach successfully retrieves the variations of the query image among a very large collection of image database, whereas the baseline returns visually similar images. Due to space limit, we will present visual results of CASIA and MIT-Adobe5K in the supplementary material.

A.5.5 Discussion

Classification Accuracy As we used the cross-entropy loss in our experiments, i.e., $\mathcal{L}_u$, presented in Section A.4, we report classification accuracy in Table A.2. As shown in the table, the classification accuracy in general slightly boosts. We believe this is because manipulated images serve as data augmentation and thus improve the off-the-shelf CNN model.

Batch Size In our implementation, we compute $\mathcal{L}_u$ over the batch instead of the entire dataset to approximately solve the formulated objective function. We consciously aware that a small batch size might leads to a larger batch normalization error and affect the approximated solution. To carefully pick a proper batch size, we experimented with three options: (1) 24 instances, (2) 80 instances, and (3) 24 instances with synchronized batch normalization⁹. We found that (2) and (3) lead to comparable results in terms of both mAP and classification accuracy while

⁸ $\alpha = 50$, $\beta_1 = 0.4$, and $\beta_2 = 0.6$

⁹i.e., synchronize the statistics of batch normalization across all GPUs

(1) results in evidently lower number. We thus believe 80 is a reasonably large batch size to approximately solve the objective function.

Distance Statistics In Section A.3.2, we analyzed the distribution of intra-instance distance, inter-instance distance, and inter-class distance on the off-the-shelf CNN feature space in Figure A.3. As we found the optimal feature space using the synthesized dataset, we plot the same statistics using the learned feature space. As shown in Figure A.7, the intra-instance distance is significantly far away from the inter-instance distance and the inter-class distance. This feature demonstrates the effectiveness of the learned feature space.

A.6 Conclusions

In this chapter, we introduce an emerging image variation retrieval problem and comprehensively analyze the problem of using off-the-shelf CNN feature on image variation retrieval. We then present a feature learning system IVAR to address image variation retrieval, which includes synthesizing the manipulated database, formulating an appropriate objective function, and solving the objective function practically using an approximation strategy of stratified sampling. Experimental results demonstrate the superior performance of the proposed approach both qualitatively and quantitatively.

Bibliography

- [1] KARG, M., A.-A. SAMADANI, R. GORBET, K. KÜHNLENZ, J. HOEY, and D. KULIĆ (2013) “Body movements for affective expression: A survey of automatic recognition and generation,” *IEEE Transactions on Affective Computing*, **4**(4), pp. 341–359.
- [2] TIAN, Y., T. KANADE, and J. F. COHN (2011) “Facial expression recognition,” in *Handbook of Face Recognition*, Springer, pp. 487–519.
- [3] ZUFFI, S., J. ROMERO, C. SCHMID, and M. J. BLACK (2013) “Estimating human pose with flowing puppets,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3312–3319.
- [4] LOPER, M., N. MAHMOOD, J. ROMERO, G. PONS-MOLL, and M. J. BLACK (2015) “SMPL: A skinned multi-person linear model,” *ACM Transactions on Graphics*, **34**(6), pp. 248:1–248:16.
- [5] YAN, S., Y. XIONG, and D. LIN (2018) “Spatial temporal graph convolutional networks for skeleton-based action recognition,” *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1), pp. 7444–7452.
- [6] KANTOROV, V. and I. LAPTEV (2014) “Efficient feature extraction, encoding and classification for action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2593–2600.
- [7] SIMONYAN, K. and A. ZISSERMAN (2014) “Two-stream convolutional networks for action recognition in videos,” in *Advances in Neural Information Processing Systems*, pp. 568–576.
- [8] CARREIRA, J. and A. ZISSERMAN (2017) “Quo vadis, action recognition? a new model and the kinetics dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 4724–4733.
- [9] WANG, L., Y. XIONG, Z. WANG, Y. QIAO, D. LIN, X. TANG, and L. VAN GOOL (2016) “Temporal segment networks: Towards good practices for deep action recognition,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 20–36.
- [10] LOPER, M., N. MAHMOOD, and M. J. BLACK (2014) “MoSh: Motion and shape capture from sparse markers,” *ACM Transactions on Graphics*, **33**(6), pp. 1–13.

- [11] CAO, Z., G. HIDALGO MARTINEZ, T. SIMON, S. WEI, and Y. A. SHEIKH (2019) “OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**(1), pp. 172–186.
- [12] KOLOTOUROS, N., G. PAVLAKOS, M. J. BLACK, and K. DANIILIDIS (2019) “Learning to reconstruct 3D human pose and shape via model-fitting in the loop,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2252–2261.
- [13] SHIFFRAR, M., M. D. KAISER, and A. CHOUCOURELOU (2011) “Seeing human movement as inherently social,” *The Science of Social Vision*, pp. 248–264.
- [14] WALLBOTT, H. G. (1998) “Bodily expression of emotion,” *European Journal of Social Psychology*, **28**(6), pp. 879–896.
- [15] MEEREN, H. K., C. C. VAN HEIJNSBERGEN, and B. DE GELDER (2005) “Rapid perceptual integration of facial expression and emotional body language,” *Proceedings of the National Academy of Sciences of the United States of America*, **102**(45), pp. 16518–16523.
- [16] DE GELDER, B. (2006) “Towards the neurobiology of emotional body language,” *Nature Reviews Neuroscience*, **7**(3), pp. 242–249.
- [17] AVIEZER, H., Y. TROPE, and A. TODOROV (2012) “Body cues, not facial expressions, discriminate between intense positive and negative emotions,” *Science*, **338**(6111), pp. 1225–1229.
- [18] LU, X., P. SURYANARAYAN, R. B. ADAMS JR, J. LI, M. G. NEWMAN, and J. Z. WANG (2012) “On shape and the computability of emotions,” in *Proceedings of the ACM International Conference on Multimedia*, ACM, pp. 229–238.
- [19] LU, X., R. B. ADAMS JR, J. LI, M. G. NEWMAN, and J. Z. WANG (2017) “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.
- [20] YE, J., J. LI, M. G. NEWMAN, R. B. ADAMS, and J. Z. WANG (2019) “Probabilistic multigraph modeling for improving the quality of crowdsourced affective data,” *IEEE Transactions on Affective Computing*, **10**(1), pp. 115–128.
- [21] KRAKOVSKY, M. (2018) “Artificial (emotional) intelligence,” *Communications of the ACM*, **61**(4), pp. 18–19.
- [22] EKMAN, P. and W. V. FRIESEN (1977) *Facial Action Coding System: A technique for the measurement of facial movement*, Consulting Psychologists Press, Stanford University, Palo Alto.

- [23] GUNES, H. and M. PICCARDI (2007) “Bi-modal emotion recognition from expressive face and body gestures,” *Journal of Network and Computer Applications*, **30**(4), pp. 1334–1345.
- [24] KLEINSMITH, A., P. R. DE SILVA, and N. BIANCHI-BERTHOUE (2006) “Cross-cultural differences in recognizing affect from body posture,” *Interacting with Computers*, **18**(6), pp. 1371–1389.
- [25] SCHINDLER, K., L. VAN GOOL, and B. DE GELDER (2008) “Recognizing emotions expressed by body pose: A biologically inspired neural model,” *Neural Networks*, **21**(9), pp. 1238–1246.
- [26] DAEL, N., M. MORTILLARO, and K. R. SCHERER (2012) “Emotion expression in body action and posture,” *Emotion*, **12**(5), pp. 1085–1101.
- [27] CAO, Z., T. SIMON, S.-E. WEI, and Y. SHEIKH (2017) “Realtime multi-person 2d pose estimation using part affinity fields,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299.
- [28] MARTINEZ, J., R. HOSSAIN, J. ROMERO, and J. J. LITTLE (2017) “A simple yet effective baseline for 3d human pose estimation,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2640–2649.
- [29] RUGGERO RONCHI, M. and P. PERONA (2017) “Benchmarking and error diagnosis in multi-instance pose estimation,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 369–378.
- [30] DARWIN, C. (1965) *The expression of the emotions in man and animals*, vol. 526, University of Chicago press.
- [31] JAMES, W. (1884) “II.—What is an emotion?” *Mind*, **9**(34), pp. 188–205.
- [32] EKMAN, P. (1992) “Are there basic emotions?” *Psychological Review*, **99**(3), pp. 550–553.
- [33] ——— (1993) “Facial expression and emotion,” *American Psychologist*, **48**(4), pp. 384–392.
- [34] EKMAN, P. and W. V. FRIESEN (1971) “Constants across cultures in the face and emotion,” *Journal of Personality and Social Psychology*, **17**(2), pp. 124–129.
- [35] MEHRABIAN, A. (1996) “Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament,” *Current Psychology*, **14**(4), pp. 261–292.
- [36] BANSE, R. and K. R. SCHERER (1996) “Acoustic profiles in vocal emotion expression,” *Journal of Personality and Social Psychology*, **70**(3), pp. 614–636.

- [37] DE GELDER, B. (2009) "Why bodies? Twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364**(1535), pp. 3475–3484.
- [38] EKMAN, P. (1965) "Differential communication of affect by head and body cues," *Journal of Personality and Social Psychology*, **2**(5), pp. 726–735.
- [39] MARTINEZ, L., V. B. FALVELLO, H. AVIEZER, and A. TODOROV (2016) "Contributions of facial expressions and body language to the rapid perception of dynamic emotions," *Cognition and Emotion*, **30**(5), pp. 939–952.
- [40] CUDDY, A. (2015) *Presence: Bringing Your Boldest Self to Your Biggest Challenges*, Hachette UK.
- [41] BURGOON, J. K., L. K. GUERRERO, and K. FLOYD (2016) *Nonverbal Communication*, Routledge.
- [42] SOKOLOV, A. A., S. KRÜGER, P. ENCK, I. KRÄGELOH-MANN, and M. A. PAVLOVA (2011) "Gender affects body language reading," *Frontiers in Psychology*, **2**, pp. 16:1–16:6.
- [43] EKMAN, P. and W. V. FRIESEN (1986) "A new pan-cultural facial expression of emotion," *Motivation and Emotion*, **10**(2), pp. 159–168.
- [44] CARMICHAEL, L., S. ROBERTS, and N. WESSELL (1937) "A study of the judgment of manual expression as presented in still and motion pictures," *The Journal of Social Psychology*, **8**(1), pp. 115–142.
- [45] KLEINSMITH, A. and N. BIANCHI-BERTHOUBE (2013) "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, **4**(1), pp. 15–33.
- [46] GENDRON, M. and L. FELDMAN BARRETT (2009) "Reconstructing the past: A century of ideas about emotion in psychology," *Emotion Review*, **1**(4), pp. 316–339.
- [47] DOUGLAS-COWIE, E., R. COWIE, I. SNEDDON, C. COX, L. LOWRY, M. MCRORIE, L. J.-C. MARTIN, J. DEVILLERS, A. ABRILIAN, S. BATLINER, ET AL. (2007) "The humane database: addressing the needs of the affective computing community," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pp. 488–500.
- [48] KLEINSMITH, A., N. BIANCHI-BERTHOUBE, and A. STEED (2011) "Automatic recognition of non-acted affective postures," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **41**(4), pp. 1027–1038.
- [49] BIRDWHISTELL, R. L. (2010) *Kinesics and context: Essays on body motion communication*, University of Pennsylvania press.

- [50] ELEFThERiADiS, S., O. RUDOVIC, and M. PANTIC (2015) “Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition,” *IEEE Transactions on Image Processing*, **24**(1), pp. 189–204.
- [51] FABIAN BENITEZ-QUIROZ, C., R. SRINIVASAN, and A. M. MARTINEZ (2016) “Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5562–5570.
- [52] XU, F., J. ZHANG, and J. Z. WANG (2017) “Microexpression identification and categorization using a facial dynamics map,” *IEEE Transactions on Affective Computing*, **8**(2), pp. 254–267.
- [53] PANTIC, M. and L. J. M. ROTHKRANTZ (2000) “Automatic analysis of facial expressions: The state of the art,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(12), pp. 1424–1445.
- [54] VIOLA, P. and M. J. JONES (2004) “Robust real-time face detection,” *International Journal of Computer Vision*, **57**(2), pp. 137–154.
- [55] WEN, Z. and T. S. HUANG (2003) “Capturing subtle facial motions in 3d face tracking,” in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, pp. 1343–1350.
- [56] LOPES, A. T., E. DE AGUIAR, A. F. DE SOUZA, and T. OLIVEIRA-SANTOS (2016) “Facial Expression Recognition with Convolutional Neural Networks: Coping with few data and the training sample order,” *Pattern Recognition*, **61**, pp. 610–628.
- [57] KAHOU, S. E., C. PAL, X. BOUTHILLIER, P. FROUMENTY, Ç. GÜLÇEHRE, R. MEMI-SEVIC, P. VINCENT, A. COURVILLE, Y. BENGIO, R. C. FERRARI, ET AL. (2013) “Combining modality specific deep neural networks for emotion recognition in video,” in *Proceedings of the ACM on International Conference on Multimodal Interaction*, ACM, pp. 543–550.
- [58] IQBAL, U., A. MILAN, and J. GALL (2017) “Posetrack: Joint multi-person pose estimation and tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2020.
- [59] GUNES, H. and M. PICCARDI (2005) “Affect recognition from face and body: early fusion vs. late fusion,” in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 3437–3443.
- [60] NICOLAOU, M. A., H. GUNES, and M. PANTIC (2011) “Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space,” *IEEE Transactions on Affective Computing*, **2**(2), pp. 92–105.

- [61] ARISTIDOU, A., P. CHARALAMBOUS, and Y. CHRYSANTHOU (2015) “Emotion analysis and classification: understanding the performers’ emotions using the LMA entities,” *Computer Graphics Forum*, **34**(6), pp. 262–276.
- [62] CORNEANU, C., F. NOROOZI, D. KAMINSKA, T. SAPINSKI, S. ESCALERA, and G. ANBARJAFARI (2018) “Survey on emotional body gesture recognition,” *IEEE Transactions on Affective Computing*.
- [63] SOOMRO, K., A. R. ZAMIR, and M. SHAH (2012) “UCF101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint arXiv:1212.0402*.
- [64] CABA HEILBRON, F., V. ESCORCIA, B. GHANEM, and J. CARLOS NIEBLES (2015) “Activitynet: A large-scale video benchmark for human activity understanding,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970.
- [65] KAY, W., J. CARREIRA, K. SIMONYAN, B. ZHANG, C. HILLIER, S. VIJAYANARASIMHAN, F. VIOLA, T. GREEN, T. BACK, P. NATSEV, ET AL. (2017) “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*.
- [66] KRIZHEVSKY, A., I. SUTSKEVER, and G. E. HINTON (2012) “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, pp. 1097–1105.
- [67] LUVIZON, D. C., D. PICARD, and H. TABIA (2018) “2d/3d pose estimation and action recognition using multitask deep learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5137–5146.
- [68] DALAL, N. and B. TRIGGS (2005) “Histograms of oriented gradients for human detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, IEEE, pp. 886–893.
- [69] FELZENSZWALB, P. F., R. B. GIRSHICK, D. MCALLESTER, and D. RAMANAN (2010) “Object detection with discriminatively trained part-based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(9), pp. 1627–1645.
- [70] YANG, Y. and D. RAMANAN (2011) “Articulated pose estimation with flexible mixtures-of-parts,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1385–1392.
- [71] PFISTER, T., J. CHARLES, and A. ZISSERMAN (2015) “Flowing convnets for human pose estimation in videos,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1913–1921.
- [72] CHARLES, J., T. PFISTER, D. MAGEE, D. HOGG, and A. ZISSERMAN (2016) “Personalizing human video pose estimation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3063–3072.

- [73] KWAK, S., M. CHO, and I. LAPTEV (2016) “Thin-Slicing for Pose: Learning to understand pose without explicit pose estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947.
- [74] ARISTIDOU, A., Q. ZENG, E. STAVRAKIS, K. YIN, D. COHEN-OR, Y. CHRYSANTHOU, and B. CHEN (2017) “Emotion control of unstructured dance movements,” in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, article 9, pp. 1–10.
- [75] VAPNIK, V. N. (1998) *Statistical learning theory*, vol. 1, Wiley New York.
- [76] EL AYADI, M., M. S. KAMEL, and F. KARRAY (2011) “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, **44**(3), pp. 572–587.
- [77] NWE, T. L., S. W. FOO, and L. C. DE SILVA (2003) “Speech emotion recognition using hidden Markov models,” *Speech Communication*, **41**(4), pp. 603–623.
- [78] TRIGEORGIS, G., F. RINGEVAL, R. BRUECKNER, E. MARCHI, M. A. NICOLAOU, S. ZAFEIRIOU, ET AL. (2016) “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 5200–5204.
- [79] PANG, B. and L. LEE (2008) “Opinion mining and sentiment analysis,” *Foundations and Trends in Information Retrieval*, **2**(1-2), pp. 1–135.
- [80] PICARD, R. W., E. VYZAS, and J. HEALEY (2001) “Toward machine emotional intelligence: Analysis of affective physiological state,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(10), pp. 1175–1191.
- [81] SOLEYMANI, M., M. PANTIC, and T. PUN (2012) “Multimodal emotion recognition in response to videos,” *IEEE transactions on Affective Computing*, **3**(2), pp. 211–223.
- [82] JOSHI, J., R. GOECKE, G. PARKER, and M. BREAKSPEAR (2013) “Can body expressions contribute to automatic depression analysis?” in *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, IEEE, pp. 1–7.
- [83] LEE, C. M. and S. S. NARAYANAN (2005) “Toward detecting emotions in spoken dialogs,” *IEEE Transactions on Speech and Audio Processing*, **13**(2), pp. 293–303.
- [84] JI, Q., P. LAN, and C. LOONEY (2006) “A probabilistic framework for modeling and real-time monitoring human fatigue,” *IEEE Transactions on Systems, Man, and Cybernetics, Part A (Systems and humans)*, **36**(5), pp. 862–875.
- [85] GUNES, H. and M. PICCARDI (2009) “Automatic temporal segment detection and affect recognition from face and body display,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **39**(1), pp. 64–84.

- [86] HALL, D. L. and J. LLINAS (1997) “An introduction to multisensor data fusion,” *Proceedings of the IEEE*, **85**(1), pp. 6–23.
- [87] SEBE, N., I. COHEN, T. GEVERS, and T. S. HUANG (2005) “Multimodal approaches for emotion recognition: a survey,” in *Electronic Imaging 2005*, International Society for Optics and Photonics, pp. 56–67.
- [88] GARG, A., V. PAVLOVIC, and J. M. REHG (2003) “Boosted learning in dynamic Bayesian networks for multimodal speaker detection,” *Proceedings of the IEEE*, **91**(9), pp. 1355–1369.
- [89] ZHENG, Y. (2015) “Methodologies for cross-domain data fusion: An overview,” *IEEE Transactions on Big Data*, **1**(1), pp. 16–34.
- [90] NGIAM, J., A. KHOSLA, M. KIM, J. NAM, H. LEE, and A. Y. NG (2011) “Multimodal deep learning,” in *Proceedings of the International Conference on Machine Learning*, pp. 689–696.
- [91] BIEL, J.-I. and D. GATICA-PEREZ (2013) “The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs,” *IEEE Transactions on Multimedia*, **15**(1), pp. 41–55.
- [92] KOSTI, R., J. M. ALVAREZ, A. RECASENS, and A. LAPEDRIZA (2017) “Emotion recognition in context,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1667–1675.
- [93] PAVLLO, D., C. FEICHTENHOFER, D. GRANGIER, and M. AULI (2019) “3D human pose estimation in video with temporal convolutions and semi-supervised training,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7753–7762.
- [94] PAVLAKOS, G., V. CHOUTAS, N. GHORBANI, T. BOLKART, A. A. OSMAN, D. TZIONAS, and M. J. BLACK (2019) “Expressive body capture: 3d hands, face, and body from a single image,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985.
- [95] NEWELL, A., K. YANG, and J. DENG (2016) “Stacked hourglass networks for human pose estimation,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 483–499.
- [96] SUN, X., J. SHANG, S. LIANG, and Y. WEI (2017) “Compositional human pose regression,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2602–2611.
- [97] KANAZAWA, A., M. J. BLACK, D. W. JACOBS, and J. MALIK (2018) “End-to-end recovery of human shape and pose,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131.

- [98] PAVLAKOS, G., L. ZHU, X. ZHOU, and K. DANILIDIS (2018) “Learning to estimate 3D human pose and shape from a single color image,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 459–468.
- [99] OMRAN, M., C. LASSNER, G. PONS-MOLL, P. GEHLER, and B. SCHIELE (2018) “Neural body fitting: Unifying deep learning and model based human pose and shape estimation,” in *Proceedings of the International Conference on 3D Vision*, IEEE, pp. 484–494.
- [100] TAN, V., I. BUDVYTIS, and R. CIPOLLA (2017) “Indirect deep structured learning for 3d human body shape and pose prediction,” in *Proceedings of the British Machine Vision Conference*, pp. 15.1–15.11.
- [101] LUO, Y., J. YE, R. B. ADAMS, J. LI, M. G. NEWMAN, and J. Z. WANG (2020) “ARBEE: Towards automated recognition of bodily expression of emotion in the wild,” *International Journal of Computer Vision*, **128**(1), pp. 1–25.
- [102] GU, C., C. SUN, D. A. ROSS, C. VONDRICK, C. PANTOFARU, Y. LI, S. VIJAYANARASIMHAN, G. TODERICI, S. RICCO, R. SUKTHANKAR, ET AL. (2018) “AVA: A video dataset of spatio-temporally localized atomic visual actions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056.
- [103] ABU-EL-HAJIA, S., N. KOTHARI, J. LEE, P. NATSEV, G. TODERICI, B. VARADARAJAN, and S. VIJAYANARASIMHAN (2016) “YouTube-8M: A large-scale video classification benchmark,” *arXiv preprint arXiv:1609.08675*.
- [104] THOMEE, B., D. A. SHAMMA, G. FRIEDLAND, B. ELIZALDE, K. NI, D. POLAND, D. BORTH, and L.-J. LI (2016) “YFCC100M: The new data in multimedia research,” *Communications of the ACM*, **59**(2), pp. 64–73.
- [105] POTAPOV, D., M. DOUZE, Z. HARCHAOUI, and C. SCHMID (2014) “Category-specific video summarization,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 540–555.
- [106] BEWLEY, A., Z. GE, L. OTT, F. RAMOS, and B. UPCROFT (2016) “Simple online and realtime tracking,” in *Proceedings of the IEEE International Conference on Image Processing*, pp. 3464–3468.
- [107] WAKABAYASHI, A., S. BARON-COHEN, S. WHEELWRIGHT, N. GOLDENFELD, J. DELANEY, D. FINE, R. SMITH, and L. WEIL (2006) “Development of short forms of the Empathy Quotient (EQ-Short) and the Systemizing Quotient (SQ-Short),” *Personality and Individual Differences*, **41**(5), pp. 929–940.
- [108] DAWID, A. P. and A. M. SKENE (1979) “Maximum likelihood estimation of observer error-rates using the EM algorithm,” *Applied Statistics*, pp. 20–28.
- [109] GWET, K. L. (2014) *Handbook of Inter-rater Reliability: The definitive guide to measuring the extent of agreement among raters*, Advanced Analytics, LLC.

- [110] RODGERS, S., L. J. KENIX, and E. THORSON (2007) “Stereotypical portrayals of emotionality in news photos,” *Mass Communication & Society*, **10**(1), pp. 119–138.
- [111] DATTA, R., D. JOSHI, J. LI, and J. Z. WANG (2006) “Studying aesthetics in photographic images using a computational approach,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 288–301.
- [112] LABAN, R. and L. ULLMANN (1971) *The Mastery of Movement*, ERIC.
- [113] KIPF, T. N. and M. WELLING (2016) “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*.
- [114] WANG, H., A. KLÄSER, C. SCHMID, and C.-L. LIU (2011) “Action recognition by dense trajectories,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3169–3176.
- [115] WANG, H. and C. SCHMID (2013) “Action recognition with improved trajectories,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551–3558.
- [116] DALAL, N., B. TRIGGS, and C. SCHMID (2006) “Human detection using oriented histograms of flow and appearance,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 428–441.
- [117] PERRONNIN, F. and C. DANCE (2007) “Fisher kernels on visual vocabularies for image categorization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.
- [118] HE, K., X. ZHANG, S. REN, and J. SUN (2016) “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
- [119] ZACH, C., T. POCK, and H. BISCHOF (2007) “A duality based approach for realtime TV-L 1 optical flow,” in *Proceedings of the Joint Pattern Recognition Symposium*, Springer, pp. 214–223.
- [120] DENG, J., W. DONG, R. SOCHER, L.-J. LI, K. LI, and L. FEI-FEI (2009) “Imagenet: A large-scale hierarchical image database,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.
- [121] POPOOLA, O. P. and K. WANG (2012) “Video-based abnormal human behavior recognition—A review,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **42**(6), pp. 865–878.
- [122] PAVLAKOS, G., X. ZHOU, K. G. DERPANIS, and K. DANIILIDIS (2017) “Coarse-to-fine volumetric prediction for single-image 3D human pose,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7025–7034.

- [123] IONESCU, C., D. PAPA, V. OLARU, and C. SMINCHISESCU (2013) “Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**(7), pp. 1325–1339.
- [124] EIGEN, D. and R. FERGUS (2015) “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2650–2658.
- [125] ALHASHIM, I. and P. WONKA (2018) “High quality monocular depth estimation via transfer learning,” *arXiv preprint arXiv:1812.11941*.
- [126] DIJK, T. V. and G. D. CROON (2019) “How do neural networks see depth in single images?” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2183–2191.
- [127] BOGO, F., A. KANAZAWA, C. LASSNER, P. GEHLER, J. ROMERO, and M. J. BLACK (2016) “Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 561–578.
- [128] MOON, G., J. Y. CHANG, and K. M. LEE (2019) “Camera distance-aware top-down approach for 3D multi-person pose estimation from a single rgb image,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10133–10142.
- [129] JOACHIMS, T. ET AL. (1999) “Transductive inference for text classification using support vector machines,” in *Proceedings of the International Conference on Machine Learning*, vol. 99, pp. 200–209.
- [130] LEE, D.-H. ET AL. (2013) “Pseudo-Label: The simple and efficient semi-supervised learning method for deep neural networks,” *Proceedings of the Workshop on Challenges in Representation Learning, in conjunction with the International Conference on Machine Learning*, **3**(2).
- [131] DE SA, V. R. (1994) “Learning classification with unlabeled data,” in *Advances in Neural Information Processing Systems*, Citeseer, pp. 112–119.
- [132] ROSENBERG, C., M. HEBERT, and H. SCHNEIDERMAN (2005) “Semi-supervised self-training of object detection models,” *In: Application of Computer Vision*, p. 29–36.
- [133] SHI, W., Y. GONG, C. DING, Z. M. TAO, and N. ZHENG (2018) “Transductive semi-supervised deep learning using min-max features,” in *Proceedings of the European Conference on Computer Vision*, pp. 299–315.
- [134] RIZVE, M. N., K. DUARTE, Y. S. RAWAT, and M. SHAH (2021) “In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning,” *arXiv preprint arXiv:2101.06329*.

- [135] FELZENSZWALB, P. F., R. B. GIRSHICK, D. MCALLESTER, and D. RAMANAN (2009) “Object detection with discriminatively trained part-based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**(9), pp. 1627–1645.
- [136] ANDRILUKA, M., L. PISHCHULIN, P. GEHLER, and B. SCHIELE (2014) “2D human pose estimation: New benchmark and state of the art analysis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3686–3693.
- [137] OSMAN, A. A., T. BOLKART, and M. J. BLACK (2020) “STAR: Sparse trained articulated human body regressor,” *arXiv preprint arXiv:2008.08535*.
- [138] WANG, H., R. A. GÜLER, I. KOKKINOS, G. PAPANDREOU, and S. ZAFEIRIOU (2020) “BLSM: A bone-level skinned model of the human mesh,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 1–17.
- [139] SAITO, S., Z. HUANG, R. NATSUME, S. MORISHIMA, A. KANAZAWA, and H. LI (2019) “PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2304–2314.
- [140] LASSNER, C., J. ROMERO, M. KIEFEL, F. BOGO, M. J. BLACK, and P. V. GEHLER (2017) “Unite the people: Closing the loop between 3d and 2d human representations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6050–6059.
- [141] MOON, G. and K. M. LEE (2020) “NeuralAnnot: Neural annotator for in-the-wild expressive 3D human pose and mesh training sets,” *arXiv preprint arXiv:2011.11232*.
- [142] ARNAB, A., C. DOERSCH, and A. ZISSERMAN (2019) “Exploiting temporal context for 3D human pose estimation in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3395–3404.
- [143] VON MARCARD, T., R. HENSCHER, M. J. BLACK, B. ROSENHAHN, and G. PONS-MOLL (2018) “Recovering accurate 3D human pose in the wild using IMUs and a moving camera,” in *Proceedings of the European Conference on Computer Vision*, pp. 601–617.
- [144] PAVLAKOS, G., X. ZHOU, and K. DANIILIDIS (2018) “Ordinal depth supervision for 3D human pose estimation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7307–7316.
- [145] GEMAN, M., S. (1987) “Statistical methods for tomographic image reconstruction,” *Bulletin of the International Statistical Institute*, **52**(4), pp. 5–21.
- [146] BESL, P. J. and N. D. MCKAY (1992) “A method for registration of 3-D shapes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **14**(2), pp. 239–256.

- [147] KATO, H., Y. USHIKU, and T. HARADA (2018) “Neural 3d mesh renderer,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3907–3916.
- [148] LIU, J., A. SHAHROUDY, M. PEREZ, G. WANG, L.-Y. DUAN, and A. C. KOT (2019) “NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**(10), pp. 2684–2701.
- [149] HE, K., G. GKIOXARI, P. DOLLÁR, and R. GIRSHICK (2017) “Mask R-CNN,” in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, pp. 2980–2988.
- [150] MEHTA, D., H. RHODIN, D. CASAS, P. FUA, O. SOTNYCHENKO, W. XU, and C. THEOBALT (2017) “Monocular 3D human pose estimation in the wild using improved cnn supervision,” in *Proceedings of the International Conference on 3D Vision*, IEEE, pp. 506–516.
- [151] JOHNSON, S. and M. EVERINGHAM (2010) “Clustered pose and nonlinear appearance models for human pose estimation,” in *Proceedings of the British Machine Vision Conference*, pp. 12.1–11.
- [152] ——— (2011) “Learning effective human pose estimation from inaccurate annotation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1465–1472.
- [153] LIN, T.-Y., M. MAIRE, S. BELONGIE, J. HAYS, P. PERONA, D. RAMANAN, P. DOLLÁR, and C. L. ZITNICK (2014) “Microsoft coco: Common objects in context,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 740–755.
- [154] TANG, S., F. TAN, K. CHENG, Z. LI, S. ZHU, and P. TAN (2019) “A neural network for detailed human depth estimation from a single image,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7750–7759.
- [155] PHILBIN, J., O. CHUM, M. ISARD, J. SIVIC, and A. ZISSERMAN (2007) “Object retrieval with large vocabularies and fast spatial matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1–8.
- [156] JEGOU, H., M. DOUZE, and C. SCHMID (2008) “Hamming embedding and weak geometric consistency for large scale image search,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 304–317.
- [157] NISTER, D. and H. STEWENIUS (2006) “Scalable recognition with a vocabulary tree,” in *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, vol. 2, IEEE, pp. 2161–2168.

- [158] BYCHKOVSKY, V., S. PARIS, E. CHAN, and F. DURAND (2011) “Learning photographic global tonal adjustment with a database of input/output image pairs,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 97–104.
- [159] LI, Y., C. FANG, J. YANG, Z. WANG, X. LU, and M.-H. YANG (2017) “Universal style transfer via feature transforms,” in *Advances in Neural Information Processing Systems*, pp. 386–396.
- [160] LI, Y., M.-Y. LIU, X. LI, M.-H. YANG, and J. KAUTZ (2018) “A closed-form solution to photorealistic image stylization,” *arXiv preprint arXiv:1802.06474*.
- [161] BROMLEY, J., I. GUYON, Y. LECUN, E. SÄCKINGER, and R. SHAH (1994) “Signature verification using a" siamese" time delay neural network,” in *Advances in neural information processing systems*, pp. 737–744.
- [162] VARIOR, R. R., M. HALOI, and G. WANG (2016) “Gated siamese convolutional neural network architecture for human re-identification,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 791–808.
- [163] SCHROFF, F., D. KALENICHENKO, and J. PHILBIN (2015) “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823.
- [164] LOWE, D. G. (2004) “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, **60**(2), pp. 91–110.
- [165] KE, Y., R. SUKTHANKAR, L. HUSTON, Y. KE, and R. SUKTHANKAR (2004) “Efficient near-duplicate detection and sub-image retrieval,” in *ACM Multimedia*, vol. 4, ACM, pp. 5–13.
- [166] BERRANI, S.-A., L. AMSALEG, and P. GROS (2003) “Robust content-based image searches for copyright protection,” in *Proceedings of the 1st ACM international workshop on Multimedia databases*, ACM, pp. 70–77.
- [167] DONG, W., Z. WANG, M. CHARIKAR, and K. LI (2012) “High-confidence near-duplicate image detection,” in *Proceedings of the ACM International Conference on Multimedia Retrieval*, ACM, pp. 1–8.
- [168] DATTA, R., D. JOSHI, J. LI, and J. Z. WANG (2008) “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Computing Surveys (Csur)*, **40**(2), pp. 1–60.
- [169] SHARIF RAZAVIAN, A., H. AZIZPOUR, J. SULLIVAN, and S. CARLSSON (2014) “CNN features off-the-shelf: an astounding baseline for recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813.
- [170] FRIDRICH, J. and J. KODOVSKY (2012) “Rich models for steganalysis of digital images,” *IEEE Transactions on Information Forensics and Security*, **7**(3), pp. 868–882.

- [171] TSAI, Y.-H., X. SHEN, Z. LIN, K. SUNKAVALLI, X. LU, and M.-H. YANG (2017) “Deep image harmonization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 3789–3797.
- [172] GOODFELLOW, I., J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, and Y. BENGIO (2014) “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, pp. 2672–2680.
- [173] KINGMA, D. P. and M. WELLING (2013) “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*.
- [174] ZHOU, P., X. HAN, V. I. MORARIU, and L. S. DAVIS (2018) “Learning rich features for image manipulation detection,” *arXiv preprint arXiv:1805.04953*.
- [175] AZULAY, A. and Y. WEISS (2018) “Why do deep convolutional networks generalize so poorly to small image transformations?” *arXiv preprint arXiv:1805.12177*.
- [176] ZHENG, S., Y. SONG, T. LEUNG, and I. GOODFELLOW (2016) “Improving the robustness of deep neural networks via stability training,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4480–4488.
- [177] NGUYEN, A., J. YOSINSKI, and J. CLUNE (2015) “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436.
- [178] JACKSON, P. T., A. ATAPOUR-ABARGHOUEI, S. BONNER, T. BRECKON, and B. OBARA (2018) “Style Augmentation: Data augmentation via style randomization,” *arXiv preprint arXiv:1809.05375*.
- [179] ZHOU, B., A. LAPEDRIZA, A. KHOSLA, A. OLIVA, and A. TORRALBA (2018) “Places: A 10 million image database for scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**(6), pp. 1452–1464.
- [180] AVIDAN, S. and A. SHAMIR (2007) “Seam carving for content-aware image resizing,” *ACM Transactions on Graphics*, **26**(3), pp. 10–es.
- [181] BABENKO, A., A. SLESAREV, A. CHIGORIN, and V. LEMPITSKY (2014) “Neural codes for image retrieval,” in *Proceedings of the European Conference on Computer Vision*, Springer, pp. 584–599.
- [182] DONG, J., W. WANG, and T. TAN (2013) “Casia image tampering detection evaluation database,” in *Proceedings of the IEEE China Summit & International Conference on Signal and Information Processing*, IEEE, pp. 422–426.

Vita

Yu Luo

Yu Luo was born in Anhui, China on March 15, 1993. He received the B.Sc. degree in Applied Physics from University of Science and Technology of China in 2015. He enrolled in the Ph.D. program in the College of Information Sciences and Technology at The Pennsylvania State University (Penn State), University Park in August 2015. He was advised by Dr. James Z. Wang and co-advised by Dr. Jia Li. His research interests include affective computing, computer vision, deep learning and human behavior understanding. He has worked as a research intern at Adobe Inc. in 2017 and 2018, and ByteDance Ltd. in 2019. In Summer 2020, he joined Amazon Lab126 as an Applied Scientist.