Int J Comput Vis manuscript No.
(will be inserted by the editor)

ARBEE: Towards Automated Recognition of Bodily Expression of Emotion In the Wild

Yu Luo · Jianbo Ye · Reginald B. Adams, Jr. · Jia Li · Michelle G. Newman · James Z. Wang

Received: date / Accepted: date

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.

Keywords Body language · emotional expression · computer vision · crowdsourcing · video analysis · perception · statistical modeling

Y. Luo (✉) · J. Ye · J.Z. Wang (✉)
College of Information Sciences and Technology,
The Pennsylvania State University, University Park, PA, USA.
E-mail: yzl5709@psu.edu

J. Ye
E-mail: yelpoo@gmail.com
Ye is currently with Amazon Lab126, Sunnyvale, CA, USA.

J.Z. Wang
E-mail: jwang@psu.edu

R.B. Adams Jr. · M.G. Newman
Department of Psychology,
The Pennsylvania State University, University Park, PA, USA.
E-mail: radams@psu.edu

M.G. Newman
E-mail: mgn1@psu.edu

J. Li
Department of Statistics,
The Pennsylvania State University, University Park, PA, USA.
E-mail: jiali@psu.edu

1 Introduction

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 (Shiffrar et al., 2011). Even at

Fig. 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.

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 (Wallbott, 1998; Meeren et al., 2005; De Gelder, 2006; Aviezer et al., 2012). 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 (Aviezer et al., 2012). 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 (Lu et al., 2012, 2017; Ye et al., 2019) from visual stimuli 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.

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 (Krakovsky, 2018). 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.

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) (Ekman and Friesen, 1977). 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 (Gunes and Piccardi, 2007; Kleinsmith et al., 2006; Schindler et al., 2008; Dael et al., 2012), which are usually of small size due to lengthy human subject study regulations. Second, bodily expression is subtle and composite. According to (Karg et al., 2013), 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 peoples behaviors, the highly cluttered background, and the often substantial differences in scale, camera perspective, and pose of the person in the frame.

In this work, we investigate the feasibility of crowdsourcing bodily expression data collection and study the computability of bodily expression using the collected data. 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** (**Bo**dy **L**anguage **D**ataset). 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.

In our work, we approach the bodily expression recognition problem with the focus of addressing the first challenge mentioned earlier. Using our proposed data collection pipeline, we have collected high quality affect annotation. With the state-of-the-art computer vision techniques, we are able to address the third challenge to a certain extent. 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 (Cao et al., 2017; Martinez et al., 2017), 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 (Ruggero Ronchi and Perona, 2017). 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.*, Martinez et al. (2017)) 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 recogni-

tion 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.

The rest of this paper is structured as follows. Section 2 reviews related work in the literature. The data collection pipeline and statistics of the BoLD dataset are introduced in Section 3. We describe our modeling processes on BoLD and demonstrate findings in Section 4, and conclude in Section 5.

2 Related Work

After first reviewing basic concepts on bodily expression and related datasets, we then discuss related work on crowdsourcing subjective affect annotation and automatic bodily expression modeling.

2.1 Bodily Expression Recognition

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 (Ekman and Friesen, 1986; Ekman, 1992), Ekman *et al.* proposed six basic emotions, *i.e.*, anger, happiness, sadness, surprise, disgust, and fear. However, as suggested by Carmichael et al. (1937) and Karg et al. (2013), 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 Mehrabian (1996), 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 (Karg et al., 2013; Kleinsmith and Bianchi-Berthouze, 2013), 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.

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 (Wallbott, 1998; Dael et al., 2012; Gunes and Piccardi, 2007; Schindler et al., 2008). Wallbott (1998) analyzes videos recorded on recruited actors and established bodily emotions as an important modality of emotion recognition. In (Douglas-Cowie et al., 2007), a human subject's emotion is elicited via interaction with computer avatar of its operator. Lu et al. (2017) crowdsourced emotion responses with image stimuli. Recently, natural or authentic emotions have generated more interest in the research community. In (Kleinsmith et al., 2011), body movements are recorded while human subjects play body movement-based video games.

Related work can be categorized based on raw data types, namely MoCap data or image/video data. For lab-setting studies such as (Kleinsmith et al., 2006, 2011; Aristidou et al., 2015), collecting motion capture data is usually feasible. Gunes and Piccardi (2007) collected a dataset with upper body movement video recorded in a studio. Other work (Gunes and Piccardi, 2007; Schindler et al., 2008; Douglas-Cowie et al., 2007) used image/video data capturing the frontal view of the poses.

Humans perceive and understand emotions from multiple modalities, such as face, body language, touch, eye contact, and vocal cues. We review the most related vision-based facial expression analysis here. 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 recent works on facial expression recognition focus on Action Unit detection, *e.g.*, (Eleftheriadis et al., 2015; Fabian Benitez-Quiroz et al., 2016). Second, the face has fewer degrees of freedom compared with the whole body (Schindler et al., 2008). To address the comparatively broader freedom of bodily movement, Karg et al. (2013) suggest the use of a movement notation system may help identify bodily expression. Other research has considered microexpressions, *e.g.*, (Xu et al., 2017), suggesting additional nuances in facial expressions. To our knowledge, no vision-based study or dataset on complete measurement of natural bodily emotions exists.

2.2 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 Ye et al. (2019), 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. Biel and Gatica-Perez (2013) 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. Kosti et al. (2017), however, used an *ad hoc* gold standard to control annotation quality and each sample in the training set was only annotated once. Lu et al. (2017) crowdsourced evoked emotions of stimuli images. Building on Lu et al. (2017), Ye et al. (2019) 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.

2.3 Automatic Modeling of 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 (Iqbal et al., 2017). 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 (Gunes and Piccardi, 2005; Nicolaou et al., 2011), 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. Gunes and Piccardi (2005) additionally defined motion protocols, such as whether the hand is facing up, and combined them with landmark displacement as features. Nicolaou et al. (2011) 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 (Kleinsmith et al., 2011, 2006; Aristidou et al., 2015; Schindler et al., 2008). 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 (Kleinsmith et al., 2011, 2006; Aristidou et al., 2015). For a more comprehensive survey of automatic modeling of bodily expression, readers are referred to the three surveys (Karg et al., 2013; Kleinsmith and Bianchi-Berthouze, 2013; Corneanu et al., 2018).

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 (Soomro et al., 2012; Caba Heilbron et al., 2015; Kay et al., 2017) and advances in deep learning (Krizhevsky et al., 2012) have accelerated the development in action recognition. To our knowledge, two-stream ConvNets-based models have been leading on this task (Simonyan and Zisserman, 2014; Wang et al., 2016; Carreira and Zisserman, 2017). 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 pose information. Yan et al. (2018), for example, modeled human skeleton sequences using a spatiotemporal graph convolutional network. Luvizon et al. (2018) 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.

3 The BoLD Dataset

In this section, we describe how we created the BoLD dataset and provide results of our statistical analysis of the data.

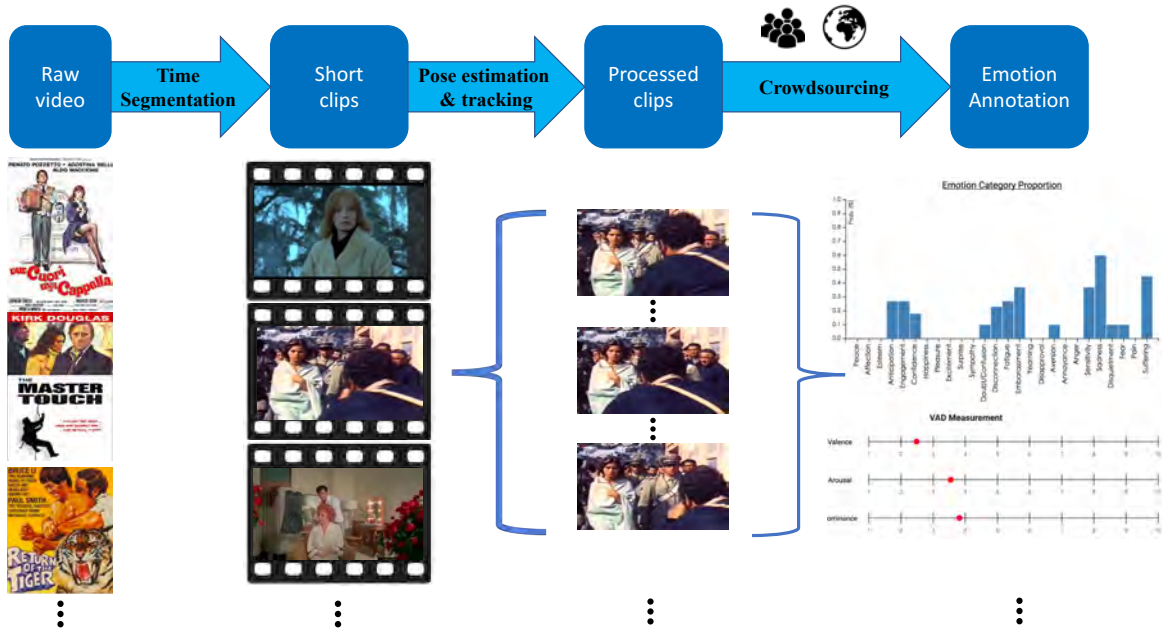


Fig. 2 Overview of our data collection pipeline. The process involves crawling movies, segmenting them into clips, estimating the poses, and emotion annotation.

3.1 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. 2 illustrates our dataset construction pipeline. We chose the movies included in a public dataset, the AVA dataset (Gu et al., 2018), 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).

3.1.1 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 (Abu-El-Haija et al., 2016; Thomee et al., 2016). In this work, we have crawled 150 movies (220 hours in total) from YouTube by the video IDs curated in the AVA dataset (Gu et al., 2018).

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 (Potapov et al., 2014). 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

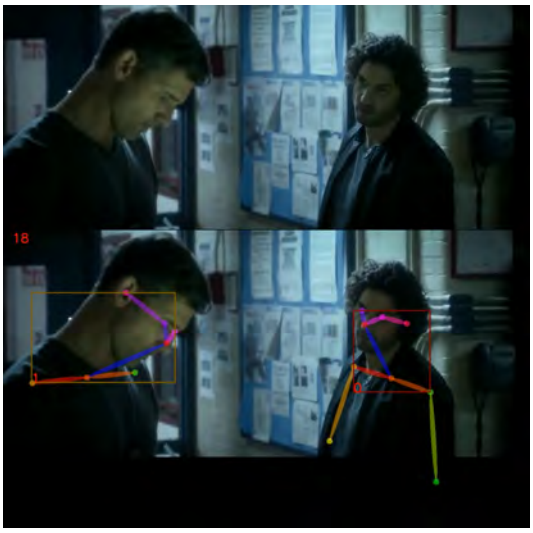


Fig. 3 A frame in a video clip, with different characters numbered with an ID (*e.g.*, 0 and 1 at the bottom left corner of red bounding boxes) and the body and/or facial landmarks detected (indicated with the stick figure).

confusion, we will use the term scene to indicate both cases.

3.1.2 Pose Estimation and Tracking

We adopted an approach to detect human body landmarks and track each character at the same time (Fig. 3). Because not all short clips contain human characters, we removed those clips without humans via pose estimation (Cao et al., 2017). Each clip was processed by a pose estimator[1] 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 (Bewley et al., 2016)[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.

3.1.3 Emotion Annotation

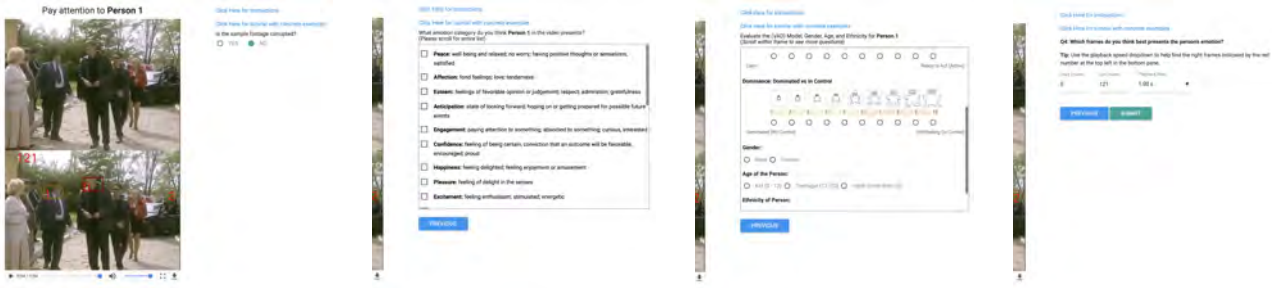
Following the above steps, we generated 122,129 short clips from these movies. We removed facial close-up clips using results from pose estimation. Concretely, we included a clip in our annotation list if the character in it has at least three visible landmarks out of the six upper-body landmarks, *i.e.*, 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 shows screenshots of our crowdsourcing website design. As a first step, participants must check if 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 (Mehrabian, 1980)). For categorical emotions, we used the list in (Kosti et al., 2017), which contains 26 categories and is a superset of the six basic emotions (Ekman, 1993). 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.

[1] `https://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

Fig. 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 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.

3.1.4 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 (Wakabayashi et al., 2006). 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 Ye et al. (2019) 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 Ye et al. (2019), 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.

3.1.5 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 (Dawid and Skene, 1979), 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 (Ye et al., 2019). 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 $i \in \{1, 2, \ldots, n\}$ and $d \in \{\text{V, A, D}\}$ in the VAD model. The final annotation is then aggregated as

$$s^d = \frac{\Sigma_{i=1}^n r_i s_i^d}{10\Sigma_{i=1}^n r_i} \ . \tag{1}$$

In the meantime, instance confidence according to the method by Ye et al. (2019) is defined as

$$c = 1 - \prod_{i=1}^{n} (1 - r_i) \ . \tag{2}$$

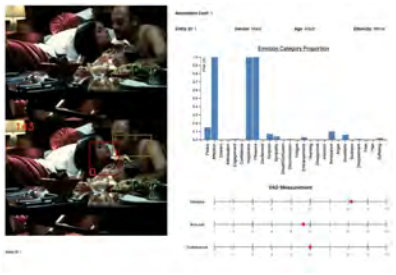
Note that we divided the final VAD score by 10 so that the data ranges between 0 and 1. Our final dataset to be used for further analysis retained only those instances with confidence higher than 0.95.

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.

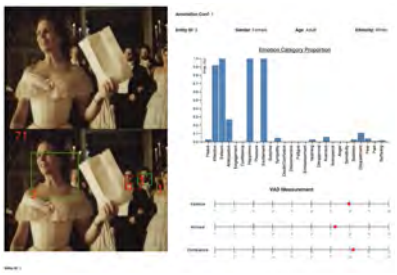
3.2 Dataset Statistics

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

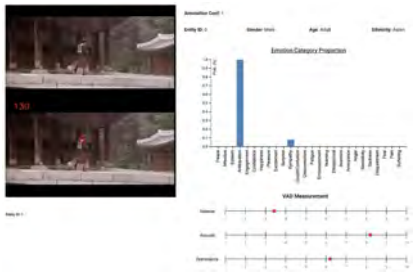
(1) peace



(2) affection

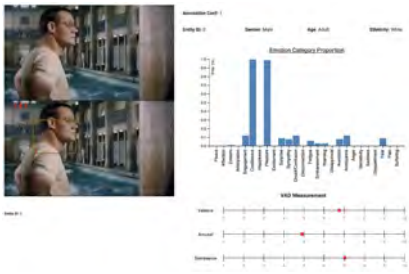


(3) esteem



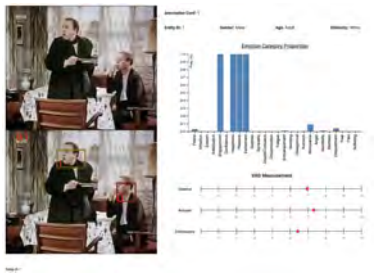
(4) anticipation

(5) engagement



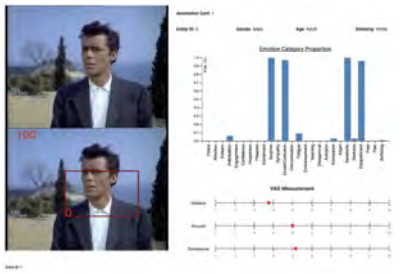
(6) confidence

(7) happiness

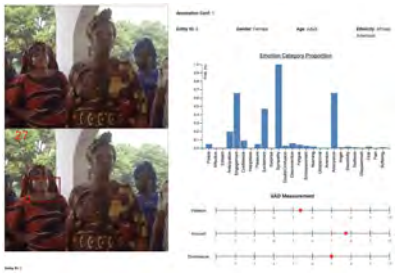


(8) pleasure

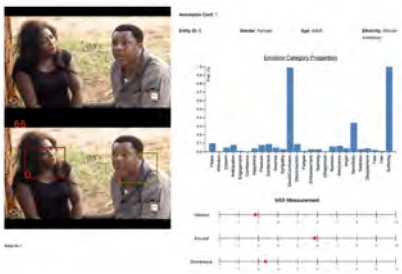
(9) excitement



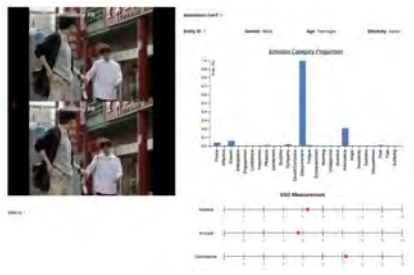
(10) surprise



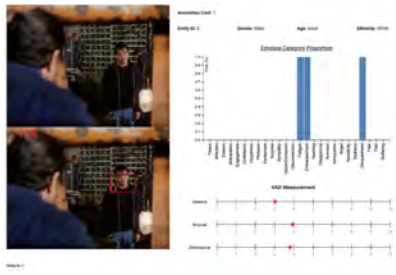
(11) sympathy



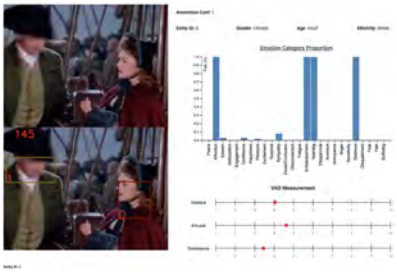
(12) doubt/confusion



(13) disconnection

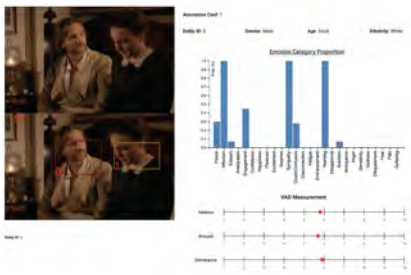


(14) fatigue

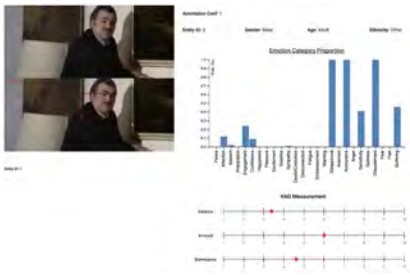


(15) embarrassment

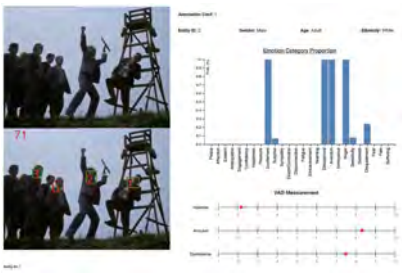
Fig. 5 Examples of high-confidence instances in BoLD for the 26 categorical emotions and two instances that were used for quality control. For each subfigure, the left side is a frame from the video, along with another copy that has the character entity IDs marked in a bounding box. The right side shows the corresponding aggregated annotation, annotation confidence c, demographics of the character, and aggregated categorical and dimensional emotion. To be continued on the next page.



(16) yearning

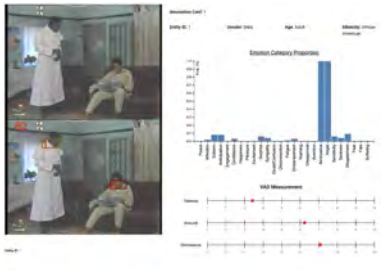


(17) disapproval

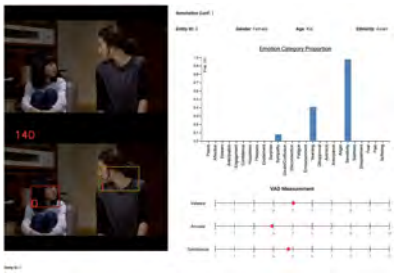


(18) aversion

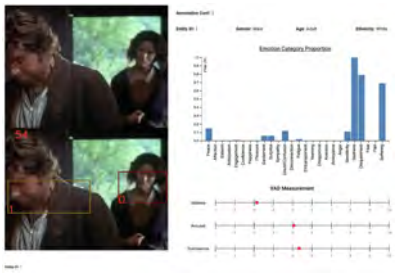
(19) annoyance



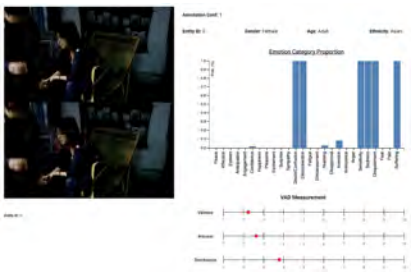
(20) anger



(21) sensitivity

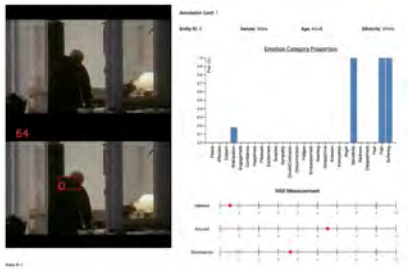


(22) sadness

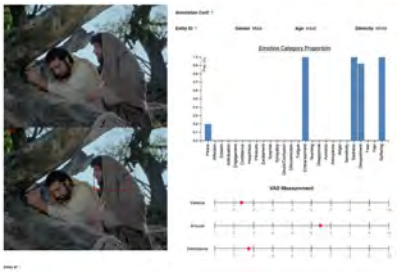


(23) disquietment

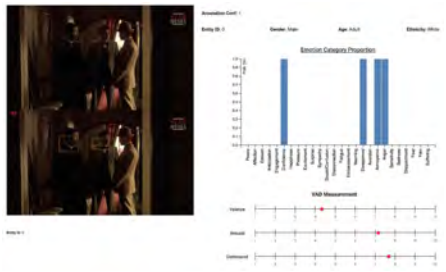
(24) fear



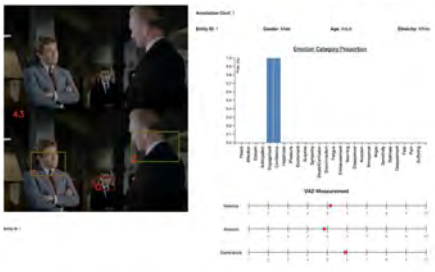
(25) pain



(26) suffering



(27) quality control



(28) quality control

Fig. 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.

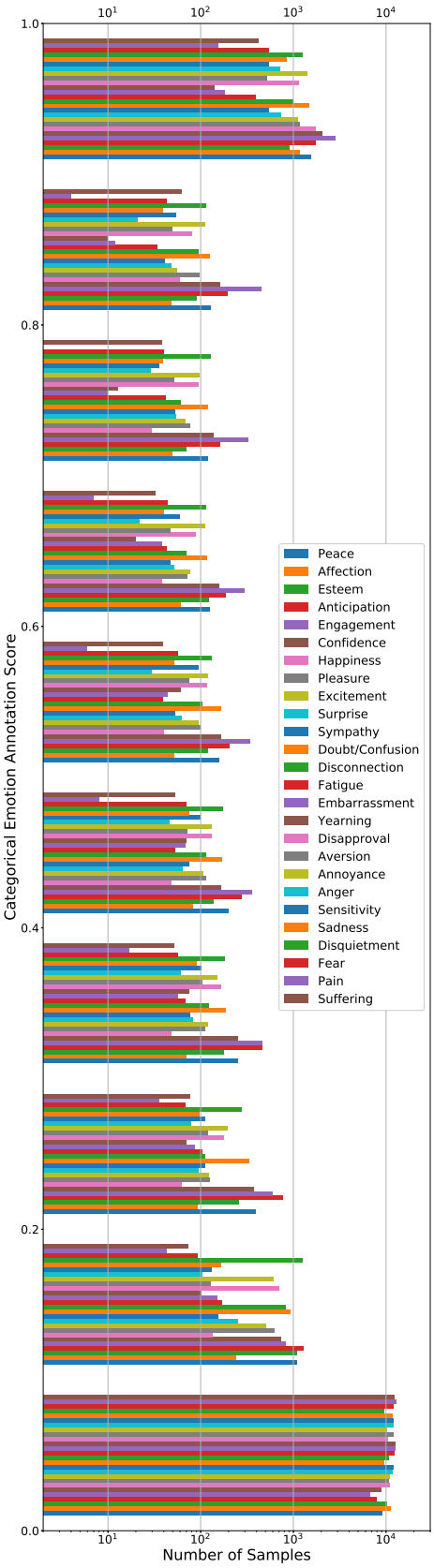


Fig. 6 Distributions of the 26 different categorical emotions.

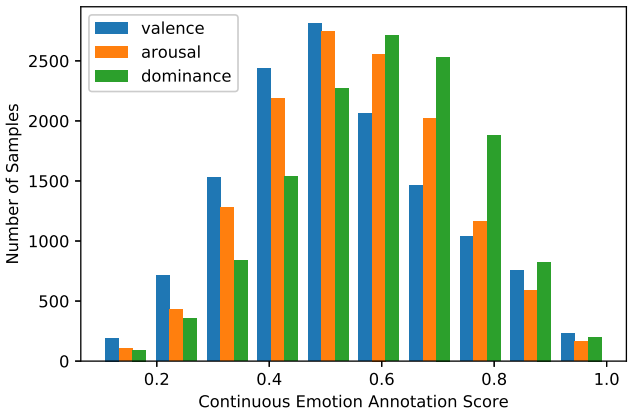


Fig. 7 Distributions of the three dimensional emotion ratings: valence, arousal, and dominance.

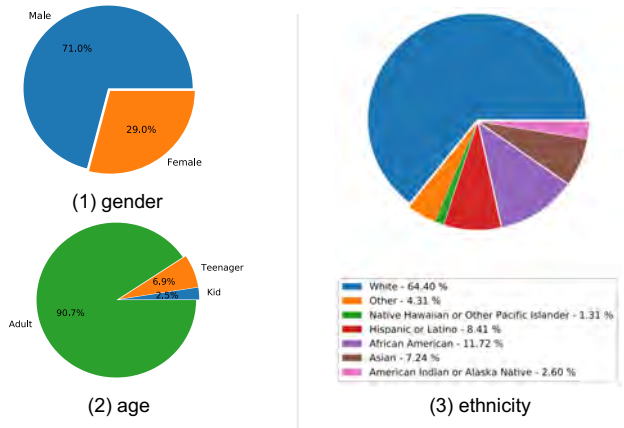


Fig. 8 Demographics of characters in our dataset.

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.

3.2.1 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. 5 shows some high-confidence instances in our dataset. Figs. 6, 7, and 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 (∼70%, 9222), the validation set (∼10%, 1153), and

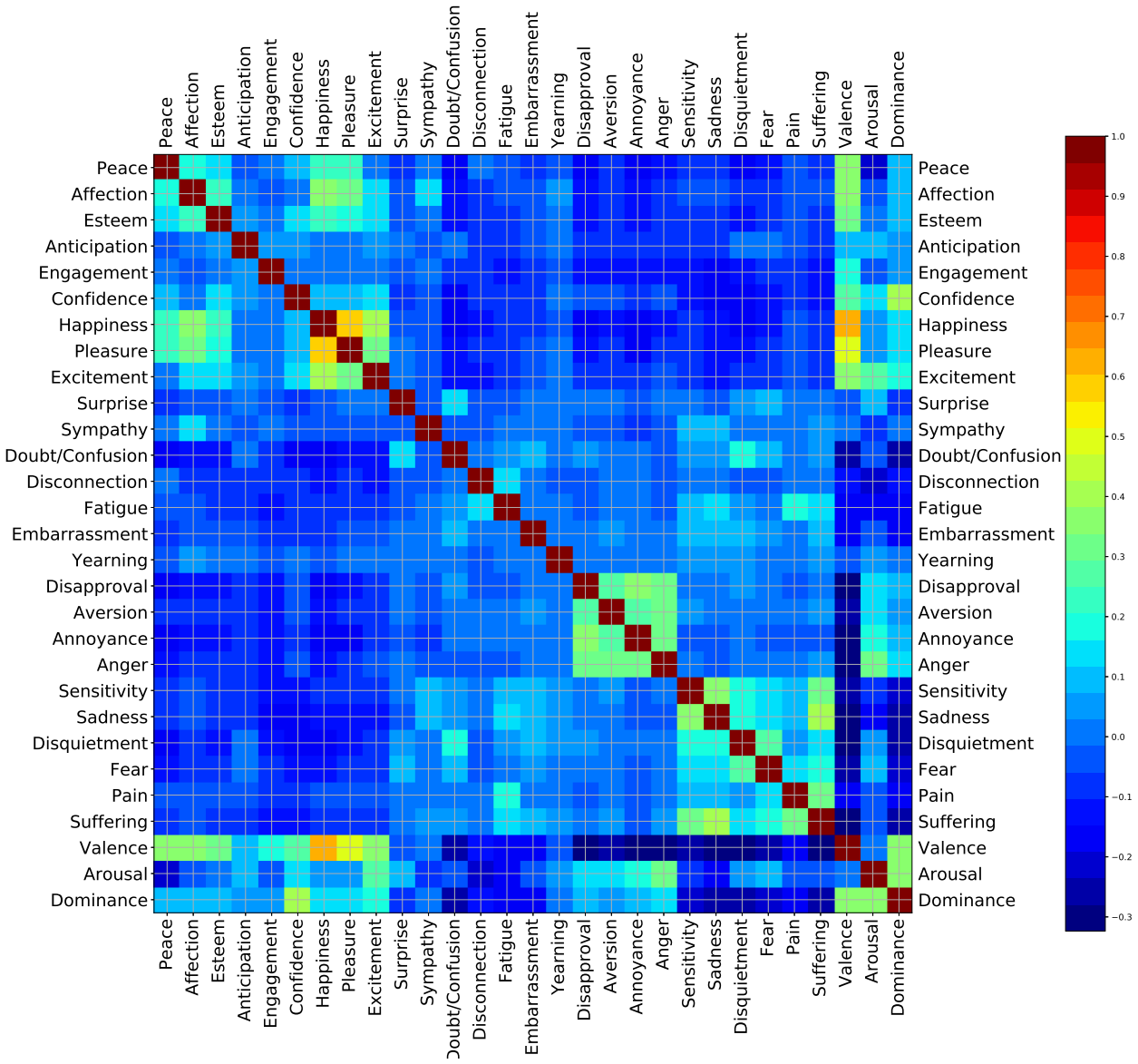


Fig. 9 Correlations between pairs of categorical or dimensional emotions, calculated based on the BoLD dataset.

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. 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 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

Table 1 Agreement among participants on categorical emotions and characters' demographic information.

Category	κ	**filtered** κ	**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

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

3.2.2 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 1 shows Fleiss' Kappa (Gwet, 2014) among participants on each categorical emotion and categorical demographic information. κ is computed on all collected annotations for each 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

$$p_j = \frac{1}{N}\sum_{i=1}^{N}\frac{n_{ij}}{n}\ , \tag{3}$$

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. (3) is revised as:

$$p_j = \frac{1}{N}\sum_{i=1}^{N}\frac{n_{ij}}{n_i}\ . \tag{4}$$

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 (Biel and Gatica-Perez, 2013). 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-instance 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.

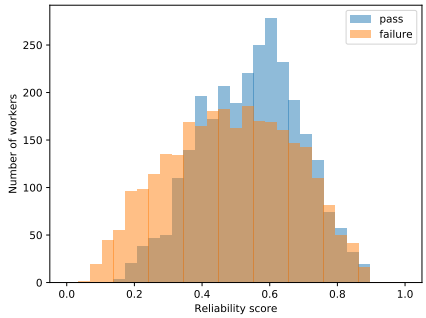


Fig. 10 Reliability score distribution among low-performance participants (failure) and non low-performance participants (pass).

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.

3.2.3 Human Performance

We explored the difference between low-performance participants and low reliability-score participants. As shown in Fig. 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 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.

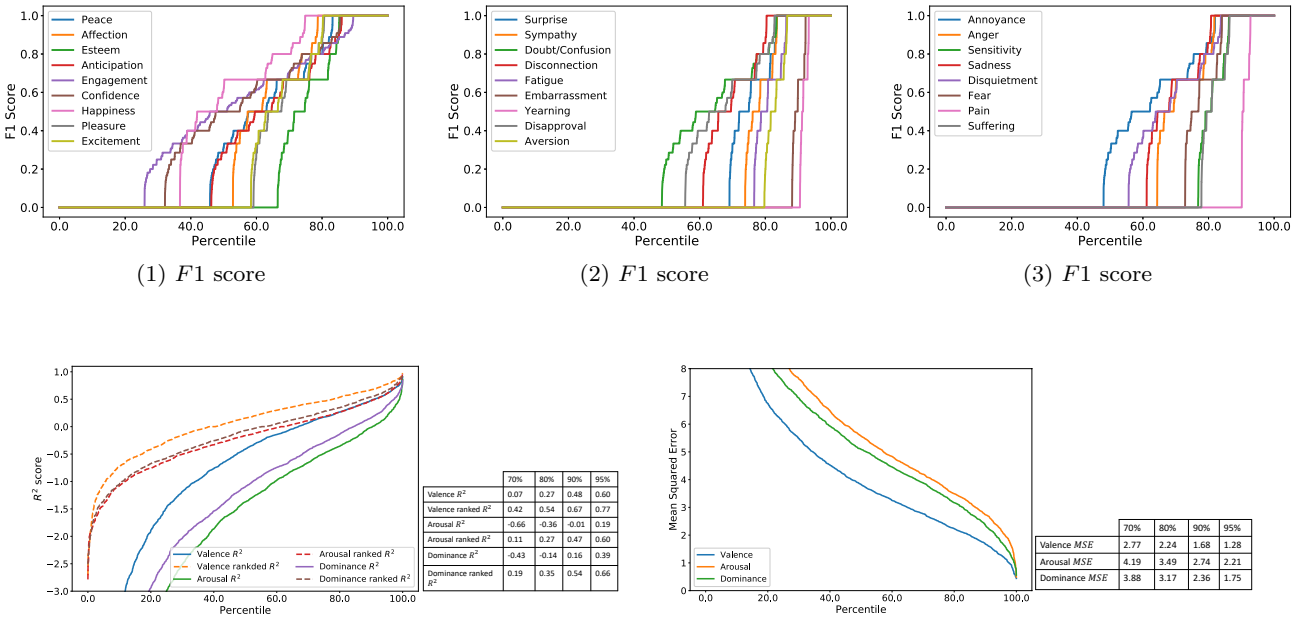
We further investigated how well humans can achieve on emotion recognition tasks. There are $5,650$ AMT participants contributing to our dataset annotation.
They represent over 100 countries (including 3,421 from the USA and 1,119 from India), with 48.4% male and 51.6% female, and an average age of 32. 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. 11 shows the cumulative distribution of participants' $F1$ scores of categorical emotions, the R^2 score, and the MSE score 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. 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.

3.2.4 Demographic Factors

Culture, gender, and age could be important factors of emotion understanding. As mentioned in Section 3.1.4, 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 cal-



(1) $F1$ score (2) $F1$ score (3) $F1$ score

	70%	80%	90%	95%
Valence R^2	0.07	0.27	0.48	0.60
Valence ranked R^2	0.42	0.54	0.67	0.77
Arousal R^2	-0.66	-0.36	-0.01	0.19
Arousal ranked R^2	0.11	0.27	0.47	0.60
Dominance R^2	-0.43	-0.14	0.16	0.39
Dominance ranked R^2	0.19	0.35	0.54	0.66

(4) $R^2 score$

	70%	80%	90%	95%
Valence MSE	2.77	2.24	1.68	1.28
Arousal MSE	4.19	3.49	2.74	2.21
Dominance MSE	3.88	3.17	2.36	1.75

(5) MSE

Fig. 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.

culated 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. 5(27-28) shows two control clips. For Fig. 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. 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.

3.2.5 Discussion

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 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 (Ye et al., 2019) 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 (Datta et al., 2006). 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.

Table 2 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

4 Bodily Expression Recognition

In this section, we investigate 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.

4.1 Learning from Skeleton

4.1.1 Laban Movement Analysis

Laban notation, originally proposed by Rudolf Laban (1971), 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 (Wallbott, 1998; Kleinsmith et al., 2006) and human computer interaction for emotion generation and classification (Aristidou et al., 2017, 2015). In our experiments, we use features listed in Table 2.

(1) natural human skeleton (2) limbs that are used in feature extraction

Fig. 12 Illustration of the human skeleton. Both red lines and black lines are considered limbs in our context.

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 (Cao et al., 2017). 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 $\nu = \{(i,j)|$ joint i and joint j are visible$\}$ be the visible set of the instance. We computed normalized pose $\hat{p}_i^t$ by

$$s = \frac{1}{T|\nu|} \sum_{(i,j)\in\nu} \sum_{t}^{T} \left\| p_i^t - p_j^t \right\|, \ \hat{p}_i^t = \frac{p_i^t}{s} \ . \tag{5}$$

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-feat-hip and right-feat-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.

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:

$$\begin{aligned} v_i^t &= \frac{\hat{p}_i^{t+\tau} - \hat{p}_i^t}{\tau} \ , \ a_i^t = \frac{v_i^{t+\tau} - v_i^t}{\tau} \ , \ j_i^t = \frac{a_i^{t+\tau} - a_i^t}{\tau} \ , \\ \hat{v}_i^t &= \left\| v_i^t \right\| \ , \ \hat{a}_i^t = \left\| a_i^t \right\| \ , \ \hat{j}_i^t = \left\| j_i^t \right\| \ . \end{aligned} \tag{6}$$

Angles, angular velocity, and angular acceleration between each pair of limbs (Fig. 12) were calculated for each pose:

$$\begin{aligned} \theta^t(i,j,m,n) &= \arccos\left(\frac{(\hat{p}_i^t - \hat{p}_j^t) \cdot (\hat{p}_m^t - \hat{p}_n^t)}{\|\hat{p}_i^t - \hat{p}_j^t\| \|\hat{p}_m^t - \hat{p}_n^t\|} \right) \ , \\ \omega_k^t(i,j,m,n) &= \frac{\theta^{t+\tau}(i,j,m,n) - \theta^t(i,j,m,n)}{\tau} \ , \\ \alpha_k^t(i,j,m,n) &= \frac{\omega^{t+\tau}(i,j,m,n) - \omega^t(i,j,m,n)}{\tau} \ . \end{aligned} \tag{7}$$

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.

4.1.2 Spatial Temporal Graph Convolutional Network

Besides handcrafted LMA features, we experimented with an end-to-end feature learning method. Following (Yan et al., 2018), 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 (Kipf and Welling, 2016) 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 (Yan et al., 2018) and trained on our task with binary cross-entropy loss and mean-squared-error loss. Our learning objective $\mathcal{L}$ can be written as:

$$\begin{aligned} \mathcal{L}_{\text{cat}} &= \sum_{i=1}^{26} y_i^{\text{cat}} \log x_i + (1 - y_i^{\text{cat}}) \log(1 - x_i^{\text{cat}}) \ , \\ \mathcal{L}_{\text{cont}} &= \sum_{i=1}^{3} (y_i^{\text{cont}} - x_i^{\text{cont}})^2 \ , \\ \mathcal{L} &= \mathcal{L}_{\text{cat}} + \mathcal{L}_{\text{cont}} \ , \end{aligned} \tag{8}$$

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.

4.2 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 (Kantorov and Laptev, 2014; Simonyan and Zisserman, 2014; Wang et al., 2016; Carreira and Zisserman, 2017). Current state-of-the-art results of activity recognition are achieved by two-stream network-based deep-learning methods (Simonyan and Zisserman, 2014). Prior to that, trajectory-based handcrafted features are shown to be efficient and robust (Wang et al., 2011; Wang and Schmid, 2013).

4.2.1 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) (Dalal et al., 2006), are widely used in activity recognition for their good performance (Wang et al., 2011; Wang and Schmid, 2013). 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 (Perronnin and Dance, 2007); 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 Kantorov and Laptev (2014) to extract trajectory-based activity features[3]. 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.

4.2.2 Deep Activity Features

Two-stream network-based deep-learning methods learn to extract features in an end-to-end fashion Simonyan and Zisserman (2014). 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. 8.

We implemented two-stream networks in PyTorch[4]. We used 101-layer ResNet as (He et al., 2016) as our network architecture. Optical flow was computed via TVL1 optical flow algorithm (Zach et al., 2007). 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 Simonyan and Zisserman (2014), 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 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) (Wang et al., 2016), 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) (Carreira and Zisserman, 2017), 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 (Wang et al., 2016; Carreira and Zisserman, 2017).

[3] `https://github.com/vadimkantorov/fastvideofeat`

[4] `http://pytorch.org/`

4.3 Results

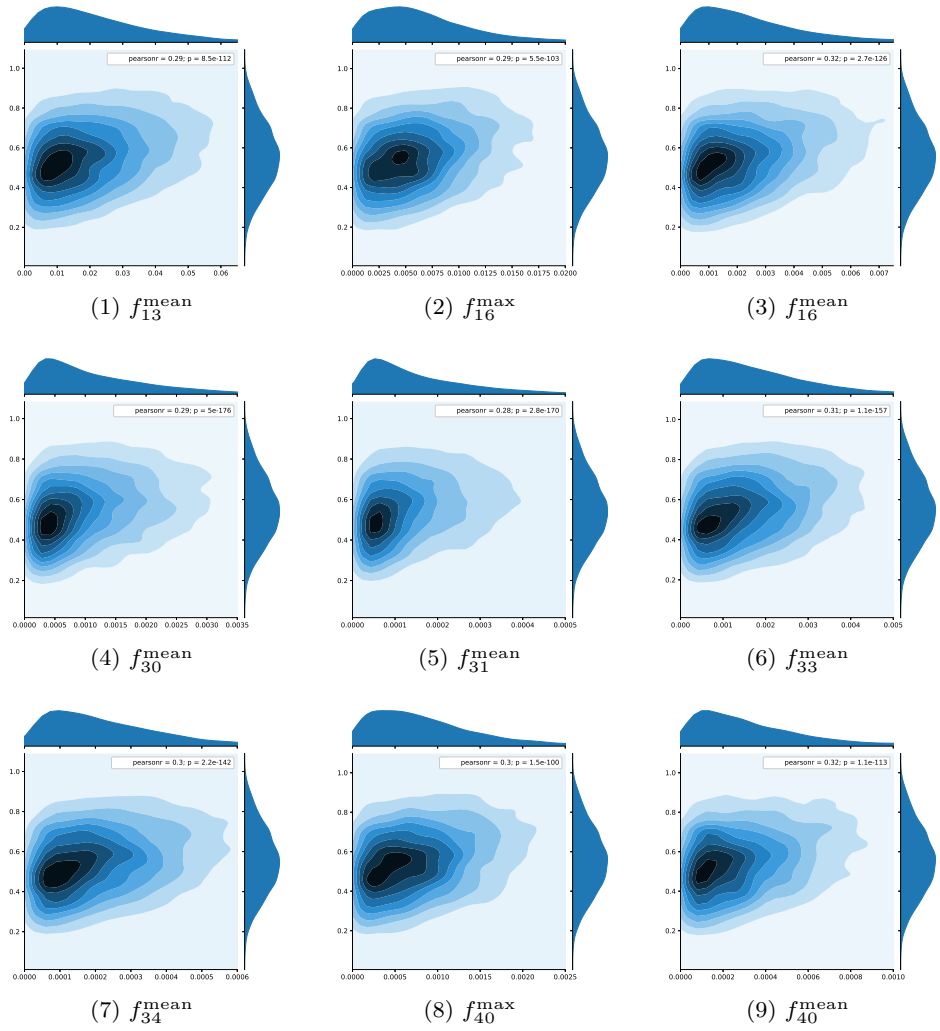
4.3.1 Evaluation Metrics

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 ($\mathrm{m}R^2$) 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:

$$\mathrm{ERS} = \frac{1}{2}\left(\mathrm{m}R^2 + \frac{1}{2}(\mathrm{mAP} + \mathrm{mRA})\right) . \tag{9}$$

4.3.2 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 2. 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. 13 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.



(1) f_{13}^{mean} (2) f_{16}^{max} (3) f_{16}^{mean}

(4) f_{30}^{mean} (5) f_{31}^{mean} (6) f_{33}^{mean}

(7) f_{34}^{mean} (8) f_{40}^{max} (9) f_{40}^{mean}

Fig. 13 Kernel density estimation plots on selected LMA features that have high correlation with arousal.

4.3.3 Model Performance

Table 3 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. 14 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. 14 and Table 1, most top-performance categories, such as affection, happiness, pleasure, excitement, sadness, anger, and pain, receive high 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. 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.

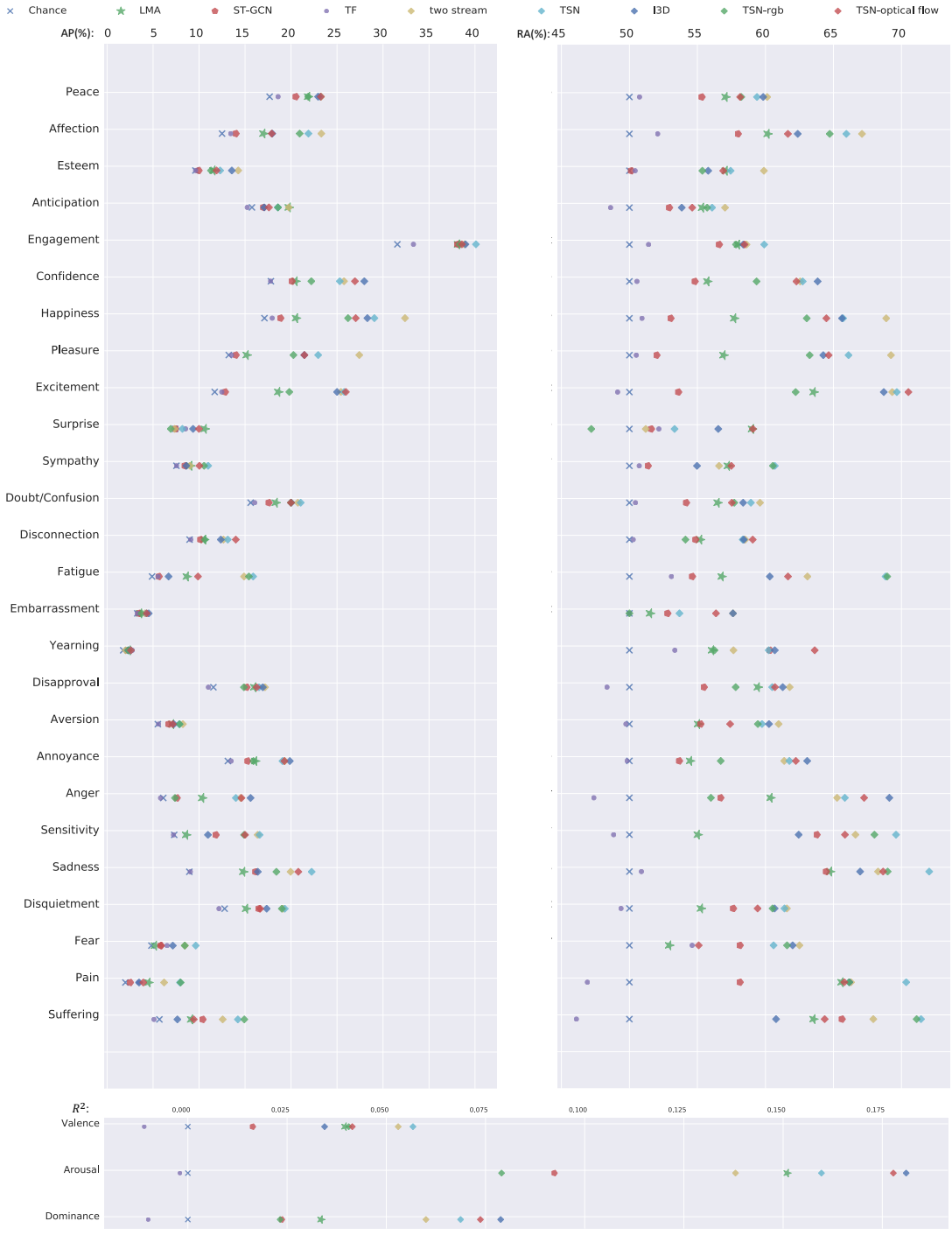


Fig. 14 Classification performance (AP: average precision on the top left, RA: ROC AUC on the top right) and regression performance (R^2 on the bottom) of different methods on each categorical and dimensional emotion.

4.4 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 (Deng et al., 2009) and action

Table 3 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 (Yan et al., 2018), TF (Kantorov and Laptev, 2014), TS-ResNet101 (Simonyan and Zisserman, 2014), I3D (Carreira and Zisserman, 2017), and TSN (Wang et al., 2016).

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

recognition model pretrained on Kinetics (Kay et al., 2017), to initialize TSN. Table 4 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 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 4 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 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 6 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

4.5 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 6 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 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 7. 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.

Table 7 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

5 Conclusions and Future Work

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.

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. 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.

Acknowledgements This material is based upon work supported in part by The Pennsylvania State University. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant No. ACI-1548562 (Towns et al., 2014). The work was also supported through a GPU gift from the NVIDIA Corporation. The authors are grateful 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. Hanjoo Kim contributed in some of the discussions. Jeremy Yuya Ong supported the data collection and visualization effort. We thank Amazon.com, Inc. for supporting the expansion of this line of research.

References

Abu-El-Haija S, Kothari N, Lee J, Natsev P, Toderici G, Varadarajan B, Vijayanarasimhan S (2016) Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:160908675

Aristidou A, Charalambous P, Chrysanthou Y (2015) Emotion analysis and classification: understanding the performers' emotions using the lma entities. Computer Graphics Forum 34(6):262–276

Aristidou A, Zeng Q, Stavrakis E, Yin K, Cohen-Or D, Chrysanthou Y, Chen B (2017) Emotion control of unstructured dance movements. In: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, article 9

Aviezer H, Trope Y, Todorov A (2012) Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science 338(6111):1225–1229

Bewley A, Ge Z, Ott L, Ramos F, Upcroft B (2016) Simple online and realtime tracking. In: Proceedings of the IEEE International Conference on Image

Processing, pp 3464–3468, DOI 10.1109/ICIP.2016.7533003

Biel JI, Gatica-Perez D (2013) The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. IEEE Transactions on Multimedia 15(1):41–55

Caba Heilbron F, Escorcia V, Ghanem B, Carlos Niebles J (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

Cao Z, Simon T, Wei SE, Sheikh Y (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

Carmichael L, Roberts S, Wessell N (1937) A study of the judgment of manual expression as presented in still and motion pictures. The Journal of Social Psychology 8(1):115–142

Carreira J, Zisserman A (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

Corneanu C, Noroozi F, Kaminska D, Sapinski T, Escalera S, Anbarjafari G (2018) Survey on emotional body gesture recognition. IEEE Transactions on Affective Computing

Dael N, Mortillaro M, Scherer KR (2012) Emotion expression in body action and posture. Emotion 12(5):1085

Dalal N, Triggs B, Schmid C (2006) Human detection using oriented histograms of flow and appearance. In: Proceedings of the European Conference on Computer Vision, Springer, pp 428–441

Datta R, Joshi D, Li J, Wang JZ (2006) Studying aesthetics in photographic images using a computational approach. In: European conference on computer vision, Springer, pp 288–301

Dawid AP, Skene AM (1979) Maximum likelihood estimation of observer error-rates using the em algorithm. Applied Statistics pp 20–28

De Gelder B (2006) Towards the neurobiology of emotional body language. Nature Reviews Neuroscience 7(3):242–249

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255

Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry L, McRorie M, Martin LJC, Devillers J, Abrilian A, Batliner S, et al. (2007) The humaine database: addressing the needs of the affective computing community. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, pp 488–500

Ekman P (1992) Are there basic emotions? Psychological Review 99(3):550–553

Ekman P (1993) Facial expression and emotion. American Psychologist 48(4):384

Ekman P, Friesen WV (1977) Facial Action Coding System: A technique for the measurement of facial movement. Consulting Psychologists Press, Stanford University, Palo Alto

Ekman P, Friesen WV (1986) A new pan-cultural facial expression of emotion. Motivation and Emotion 10(2):159–168

Eleftheriadis S, Rudovic O, Pantic M (2015) Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. IEEE Transactions on Image Processing 24(1):189–204

Fabian Benitez-Quiroz C, Srinivasan R, Martinez AM (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

Gu C, Sun C, Ross DA, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R, 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

Gunes H, Piccardi M (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

Gunes H, Piccardi M (2007) Bi-modal emotion recognition from expressive face and body gestures. Journal of Network and Computer Applications 30(4):1334–1345

Gwet KL (2014) Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Advanced Analytics, LLC

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

Iqbal U, Milan A, Gall J (2017) Posetrack: Joint multi-person pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2011–2020

Kantorov V, Laptev I (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

Karg M, Samadani AA, Gorbet R, Kühnlenz K, Hoey J, Kulić D (2013) Body movements for affective expression: A survey of automatic recognition and generation. IEEE Transactions on Affective Computing 4(4):341–359

Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, et al. (2017) The kinetics human action video dataset. arXiv preprint arXiv:170506950

Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907

Kleinsmith A, Bianchi-Berthouze N (2013) Affective body expression perception and recognition: A survey. IEEE Transactions on Affective Computing 4(1):15–33

Kleinsmith A, De Silva PR, Bianchi-Berthouze N (2006) Cross-cultural differences in recognizing affect from body posture. Interacting with Computers 18(6):1371–1389

Kleinsmith A, Bianchi-Berthouze N, Steed A (2011) Automatic recognition of non-acted affective postures. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41(4):1027–1038

Kosti R, Alvarez JM, Recasens A, Lapedriza A (2017) Emotion recognition in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1667–1675

Krakovsky M (2018) Artificial (emotional) intelligence. Communications of the ACM 61(4):18–19

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp 1097–1105

Laban R, Ullmann L (1971) The Mastery of Movement. ERIC

Lu X, Suryanarayan P, Adams Jr RB, Li J, Newman MG, Wang JZ (2012) On shape and the computability of emotions. In: Proceedings of the ACM International Conference on Multimedia, ACM, pp 229–238

Lu X, Adams Jr RB, Li J, Newman MG, Wang JZ (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

Luvizon DC, Picard D, Tabia H (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

Martinez J, Hossain R, Romero J, Little JJ (2017) A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2640–2649

Meeren HK, van Heijnsbergen CC, de Gelder B (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):16518–16523

Mehrabian A (1980) Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies. The MIT Press, Cambridge

Mehrabian A (1996) Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology 14(4):261–292

Nicolaou MA, Gunes H, Pantic M (2011) Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Transactions on Affective Computing 2(2):92–105

Perronnin F, Dance C (2007) Fisher kernels on visual vocabularies for image categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1–8

Potapov D, Douze M, Harchaoui Z, Schmid C (2014) Category-specific video summarization. In: European conference on computer vision, Springer, pp 540–555

Ruggero Ronchi M, Perona P (2017) Benchmarking and error diagnosis in multi-instance pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 369–378

Schindler K, Van Gool L, de Gelder B (2008) Recognizing emotions expressed by body pose: A biologically inspired neural model. Neural Networks 21(9):1238–1246

Shiffrar M, Kaiser MD, Chouchourelou A (2011) Seeing human movement as inherently social. The Science of Social Vision pp 248–264

Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp 568–576

Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:12120402

Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ (2016) Yfcc100m: The new data in multimedia research. Communications of the ACM 59(2):64–73

Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S, Lifka D, Peterson GD, et al. (2014) Xsede: accelerating scien-

tific discovery. Computing in Science & Engineering 16(5):62–74

Wakabayashi A, Baron-Cohen S, Wheelwright S, Goldenfeld N, Delaney J, Fine D, Smith R, Weil L (2006) Development of short forms of the empathy quotient (eq-short) and the systemizing quotient (sq-short). Personality and Individual Differences 41(5):929–940

Wallbott HG (1998) Bodily expression of emotion. European Journal of Social Psychology 28(6):879–896

Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3551–3558

Wang H, Kläser A, Schmid C, Liu CL (2011) Action recognition by dense trajectories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3169–3176

Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision, Springer, pp 20–36

Xu F, Zhang J, Wang JZ (2017) Microexpression identification and categorization using a facial dynamics map. IEEE Transactions on Affective Computing 8(2):254–267

Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence

Ye J, Li J, Newman MG, Adams RB, Wang JZ (2019) Probabilistic multigraph modeling for improving the quality of crowdsourced affective data. IEEE Transactions on Affective Computing 10(1):115–128

Zach C, Pock T, Bischof H (2007) A duality based approach for realtime tv-l 1 optical flow. In: Proceedings of the Joint Pattern Recognition Symposium, Springer, pp 214–223