Panel
Bodily Expressed Emotion Understanding Research: A Multidisciplinary Perspective

James Z. Wang (Chair)[1][0000−0003−4379−4173], Norman Badler[2][0000−0001−8659−3828], Nadia Berthouze[3][0000−0001−8921−0044], Rick O. Gilmore[1][0000−0002−7676−3982], Kerri L. Johnson[4], Agata Lapedriza[5,6][0000−0002−5248−0443], Xin Lu[7], and Nikolaus Troje[8][0000−0002−1533−2847]

[1] The Pennsylvania State University, University Park, Pennsylvania, USA
[2] University of Pennsylvania, Philadelphia, Pennsylvania, USA
[3] University College London, London, United Kingdom
[4] University of California, Los Angeles, California, USA
[5] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[6] Universitat Oberta de Catalunya, Barcelona, Spain
[7] Adobe Inc., San Jose, California, USA
[8] York University, Toronto, Ontario, Canada

Abstract. Developing computational methods for bodily expressed emotion understanding can benefit from knowledge and approaches of multiple fields, including computer vision, robotics, psychology/psychiatry, graphics, data mining, machine learning, and movement analysis. The panel, consisting of active researchers in some closely-related fields, attempts to open a discussion on the future of this new and exciting research area. This paper documents the opinions expressed by the individual panelists.

1 Introduction

Bodily expressed emotion understanding is a complex and highly challenging research problem that requires researchers to use knowledge and approaches from some quite distinct fields. For instance, computer vision, robotics, psychology/psychiatry, graphics, data mining, machine learning, and movement analysis can all play a critical role in solving this problem. Researchers from individual fields have developed theories, techniques, and systems to address this problem. As larger and larger datasets are being collected and shared with the broad research community (*e.g.*, the BoLD dataset made available at the First International Workshop on Bodily Expressed Emotion Understanding (BEEU) [6]), it is becoming important for researchers from different fields to join forces and tackle the problem together.

The goal of this multidisciplinary panel is to open a discussion on how various fields can work together in the current data-driven research environment.

Due to the COVID-19 pandemic, the panel was organized as a written panel. Five questions were sent to the panelists a few weeks in advance. Each panelist independently came up with a written response. Recorded in the remaining sections are the opinions of the individual panelists. When the written comments were collected, no length limitation was set. In the interest of keeping a record, the panel chair's personal opinions are also incorporated into this article.

2 Datasets and Benchmarks

What new kind of datasets and/or benchmarks can be helpful in bodily expressed emotion understanding research?

Norman Badler: As we discovered in our work with the effect of personality on motion performance, people can make personality type assessments from observing behaviors, especially when they can compare the same motion between differing personality types. Accordingly, it may be important to understand a subject's baseline personality characteristics (say from the 5-factor OCEAN model) as a tag on acquired motion datasets. Human motions are varied enough that factoring out personality influences may give more accurate base motions from which short-duration behaviors (such as emotional expression) may be more reliably extracted.

Similar observations apply to facial affect. Our work on FacEMOTE led us to believe that some facial action behavior variation could be described globally (per subject) [1]. Combined with the personality work, this implies that facial animation and affect understanding could be aided by an understanding of the subject's personality type. Such personality type annotations are not normally collected in motion datasets.

Nadia Berthouze:

- Naturalistic datasets: understanding emotions expressed through the body as people engaged in their everyday activity
- Situated multimodal datasets to understand affective body in depth and in the context of other modalities and situated interactions
- We need benchmark datasets that help to work across sensor types. Two worlds: computer vision and wearable-based research
- We need datasets across a variety of applications to ensure approaches can be tested for their ability to generalize to different contexts
- It's important that used ethical and GDPR (or related rules) procedure are clearly available with the datasets and consider these issues broadly to ensure that the researchers around the world can download the datasets and use them. At the moment in the UK we are not allow to download benchmark datasets if not compliant with GDPR, or need special permission
- We are missing work on affective hand gestures or touch as an extension of gesture

Rick O. Gilmore: Researchers should commit to and support the creation of an open database of images and videos that can be used to share training and test sets and labelled data in common, interoperable data formats. The database should support API queries so that reproducible computational workflows can be achieved. Researchers who draw upon the database should commit to uploading their results so that the community can accumulate sets of model fits to the same data which can then be compared and contrasted.

Databrary.org, the world's only restricted access data library specialized for storing, streaming, and sharing video could be adapted or extended for this purpose. Databrary has developed a policy framework for sharing identifiable data with participant permission that could also be extended. Indeed, concerns about personal privacy will continue to grow, so I think that collaboration with data repositories that have some of these problems in hand will become increasingly important.

Kerri L. Johnson: One of the challenges for this research is that for many researchers, particularly those who are not at R1 universities[9], the necessary equipment is cost and space prohibitive. Without sizable funding through either equipment grants or start-up funds, many simply cannot conduct research in this area. This makes the publication of open databases of body movements, both in raw and visual formats, all the more important. Unfortunately, many of the existing stimulus archives are insufficient for answering anything but very narrow questions. Expanding those archives to include more variety of movements and more diversity of individuals will expand the research questions that can be probed.

Agata Lapedriza: It depends on the specific problem that one wants to study. For example, to create systems for detecting gestures or body postures that correlate with the intended emotion expression, movies can be an interesting data source. Actors in movies behave in a way that makes it easy for us to understand what they are communicating. That means that, from movies, we should be able of learning some common gestures or body poses that people make for communicating emotions. However, if the goal is to learn a personalized model for recognizing how a specific individual expresses emotion, then it will be necessary to collect data of the specific individual, ideally with self-report emotion expression labels.

Xin Lu: In general, new datasets/benchmarks can be helpful by considering the following aspects: (1) Datasets should include a rich collection of continuous body movements together with contextual information, such as short/long videos with event tags (*e.g.*, cooking, working, delivering a speech or watching movies) or textural descriptions. The camera position (or a relative position of the camera to the person) should be included as metadata of the collected videos. (2) Datasets

[9] R1: Doctoral Universities – Very high research activity, as classified in the Carnegie Classification of Institutions of Higher Education.

should include body expressions collected from diverse groups of individuals. Demographic information is helpful for later analyses of body expressions in different age group, gender, and geographic region. (3) To make a dataset useful practically, having instance segmentation masks and information of each instance is crucial.

Creating new datasets by constraining the camera location and the event is another direction to study bodily expressed emotion in context. For instance, video data can be collected by installing a couple of cameras in fixed positions in a few conference rooms, classrooms, gyms, or landmarks.

Besides, body expression closely impact applications in health care and physical therapy. For instance, body expression may indicate a patient's emotional status, such as nervousness, anxiety, and excitement, which may help doctors diagnose and proceed with proper treatments. Also, body expression can help understand the sitting or sleeping posture in daily life, and bad posture could potentially be identified. Changes of body expressions of an individual in a day or a week may also indicate one's emotional status, such as tiredness or elation.

Nikolaus Troje: Depends very much on what we want to learn:

Ideally, we want data bases where we have detailed body kinematics of the individual, on the one hand, and information about the "true" emotional state of that individual, on the other hand. In order to subject such data to end-to-end machine learning, we need large amounts of such data. In order to ensure ecological significance, they should come from naturalistic settings. That way we could learn generative models that can serve both analysis and synthesis.

That would be the ideal scenario, but it is hardly practical. However, there are a number of strategies and workarounds that could help to get close to that point.

On the stimulus side, we may be able to replace the "ground truth" we could obtain with accurate motion capture (MoCap) technology, with information obtained from natural, markerless video. However, that requires better models to be used for markerless pose estimation. Learning these, requires yet other kind of training data and therefore different databases. Here, we need calibrated video, on the one hand, and "ground truth"-delivering motion capture data, on the other hand. However these can come from existing databases and do not necessarily require annotations in terms of emotion. bmlMoVi [3] is a database that has the potential to help learing such models. The large amount of motion capture data merged into AMASS [7] and that fact that it is fully compatible with SMLP [5], a generative model that can reconstruct fully skinned individual body shapes from motion capture data, could also be used to generate large training datasets that connect motion capture data with computer-graphically simulated video data.

On the annotation side, we also may have to make compromises. The "true" emotion of a person is hard to assess. Emotion induction is possible but also tedious and error-prone (*e.g.* [8]). For that reason, the annotation of true emotion of an individual is often replaced with ratings on how this person is perceived.

Here, observers watch a video or other visual material and then rate the assumed emotional state of the presented person. Such data can be obtained with crowd sourcing methods. While they do not necessarily speak to the true emotion of the person, they can provide accurate measures as to how a person is being perceived. Wherever a distinction between "true" and perceived emotion is critical, the relation between the two needs to be evaluated in separate experiments which then would require their own databases.

James Z. Wang: Although we have developed and released the BoLD dataset, which is based on videos slips from movies [6], we believe there are at least several ways to substantially expand the data collection effort for bodily expressed emotion understanding.

First, besides collecting emotion annotation data based on the normal or typical population, it would be important to collect a human behavior and interaction dataset based on the *atypical population* (*e.g.*, patients visiting a psychological clinic).

Second, to help bridge the enormous gap between computable human pose and movement information and bodily expressed emotions, it would be helpful to build a middle layer of data with movement annotations. The Laban Movement Analysis can be a valuable way to annotate human movements.

Third, a comprehensive dataset to study emotion should have a significant component that integrates both multi-camera video, depth information, and MoCap data so that more accurate human pose modeling can be possible.

Fourth, a large-scale dataset for machine learning and statistical modeling needs to be relatively balanced in terms of subjects' demographics (*e.g.*, gender, age, race, and ethnic groups) to reduce the potential biases in the learning process.

Finally, a successful dataset should provide user services, such as access and search API and benchmarking, to facilitate multidisciplinary research and stimulate innovation.

3 Challenge Problems

What challenge problems can potentially be important to academia and/or the industry?

Norman Badler: Human performance is multi-layered. An emotional display within an individual may vary due to mood, context, personality, or culture. For example, angry gestures may be subdued in more publicly reverent environments or accentuated in large scale competitive situations. To me, the challenge is managing the layering through explicit representations rather than trying an all-in-one leap via machine learning, perhaps, from observed behavior to hypothesized internal causes or motivation. Such variations are known to exist across cultures, making it crucial that industrial-strength machine vision systems that are being used to watch people make behavioral judgements in the context of

who is being observed and what the baselines are for nominal individuals of the community. On the other hand, real stereotypes exist and need to be carefully and deliberately removed from automated assessments that trigger moral or legal actions.

Nadia Berthouze:

- Real-world applications
- Ethical issues
- Working across sensors for ubiquitous leverage of body expressions

Rick O. Gilmore: I think the problem of accurately detecting emotions "in the wild" – from videos collected in every day settings and from possibly mobile cameras – is hard and interesting.

Kerri L. Johnson: To my thinking, a solid appreciation of how dynamic interactions impact social perception lags woefully behind our understanding of face perception. Part of that is undoubtedly due to the costs to entry in this area, relative to face perception research. Increasing accessibility to raw data will help in this regard.

Agata Lapedriza: Automatic emotion recognition in the wild is still immature. There are a lot of problems and challenges that are interesting for academic research, like the ones that I mention below. For the industry, and for society more generally, one of the main challenges is to make sure that emotion recognition software is used correctly and fairly, and just for those applications that benefit humans.

Xin Lu: While we tend to think about patterns across a large group of populations in terms of bodily expressed emotion, a potentially more interesting problem is to analyze bodily expressed emotion of an individual.

People are likely to be more relaxed and share more bodily expressed emotion unconsciously in a private space. Meanwhile, body expression is likely to be richer comparing with publicly shared body expressions. However, data out of those scenarios is usually private. How can we analyze these data assuming we cannot see most of these data due to privacy concerns? If we can, how can individuals benefit from these data? How can we analyze similarities and differences of these data across individuals? How would those patterns benefit the society? I believe this is an interesting direction to explore, which involves sufficient technical challenges and application potentials.

Nikolaus Troje: The distinction between veridical and faked emotion is an interesting topic in itself. Bodily expressions of emotions can be faked, but only to a certain degree. Fake smile has been characterized relatively well and computer vision algorithms have been designed to discriminate fake from veridical smile. However, no database or benchmark framework exists that could be used to systematically test and compare existing algorithms. Extending research to other

emotions (in addition to smile) may help to isolate invariants that discriminate between veridical and fake emotional expression in more general terms.

James Z. Wang: Comprehensive understanding of human's bodily expressed emotions has wide-ranging implications. Among others, there are clear applications in healthcare, commerce, robotics, and public safety. I believe computer vision and machine learning researchers need to partner with psychologists and key industrial players to define challenge questions that are both valid from a psychological perspective and meaningful or potentially impactful in the real world. And because real-world applications in this area will have impact to people's lives, it is critical to develop explainable, interpretable, and fair machine learning approaches.

4 Needed Breakthroughs

What theoretical, methodological, or technological breakthroughs are needed?

Norman Badler: Here I am old-fashioned: I like to know what my representation does and doesn't do, how the representation processes and interacts with real data, and what the output states or parameters mean. This goes along with my belief in layering representations to understand interactions, rather than lumping all decisions into a single complex machine learning (ML) system. I think there is a role for ML to develop such smaller steps. Overall the result may be a series of interacting ML systems, but we can look in-between them and gain some useful understanding of what is going on (or dare I say what they are "thinking").

Nadia Berthouze:

- HAR and Affective computing fields should get together. Continuous detection of affect across activities
- Studying affective body expressions as situated interactions rather than based on acontextual models
- Going beyond what we see: looking at physiological and neural mechanisms that drive movements
- Network architectures designed for capturing movement

Rick O. Gilmore: Most computer vision models ignore time, and yet human behaviors and especially bodily expressed emotions are temporally bound phenomena. We need theories that incorporate dynamics and methodological and technological breakthroughs that make this practical.

We need more research that compares model performance to human-labelled ground truth but also models against one another. We also need research with human coders in the loop to determine where specific models fail and why they fail. It may also be useful to compare weighted "mixtures" of models, especially

if those are trained on different datasets. This may be an inexpensive way to work toward models that generalize.

Finally, research in this area needs to embrace "multi-modality." By this I mean that humans and non-human animals express emotions through facial expressions, body postures, actions, and vocalizations. Researchers who figure out how to combine these sources will likely create systems that perform better than those that do not embrace a multimodal approach.

Kerri L. Johnson: At present, research in this area remains highly segmented, with vision, computational, and social scientists largely pursuing their work isolated from one another. Consequently, the insights and discoveries can be unnecessarily narrow. While some, including the members of this panel, have sought to bridge the vision, computational, and social approaches, doing so remains rare. It's at this nexus where exciting and groundbreaking discoveries will be made.

Agata Lapedriza: Here are two examples:

- Understanding the context and integrating the analysis of context to the recognition pipeline. A facial expression or a gesture can mean or communicate different messages, depending on the context.
- Dealing with subjectivity: emotion perception is subjective. Annotations on emotions provided by annotators are subjective. This means that the same input can have different labels that are correct, since given a specific input, different people might perceive different emotions and all the perceptions are correct. This is something less common in object or scene categorization, for instance. If an object is a car it is not a bicycle. However, a person can look sad to someone and angry to someone else. And both perceptions are correct, because they are just perceptions. Modelling subjectivity requires learning methods that can correctly deal with different degrees of agreement, different opinions, and different degrees of uncertainty.

Xin Lu: In the past, when we lacked data, we developed methodologies and technologies to overcome the data scarceness. In recent years, when we were equipped with a sufficient amount of data together with computational power, we identified useful patterns and used learned knowledge (usually represented by models) to automate repetitive workflows and alleviate cognitive load for individuals. In the coming future, as the amount of data is soaring, one model may not effectively encode all the information in a complex problem. I believe breakthroughs are needed in the below two aspects: 1. To be able to identify patterns selectively according to context. 2. To be able to incrementally enrich the model as data is accumulating without forgetting existing info.

Nikolaus Troje: Reliable characterization of emotion from non-facial features has interesting applications in situations where the assessment of facial features is difficult because users wear head gear such as shutter glasses or HMDs. Avatars of the players of computer games or users of telecommunication and teleconferencing systems that make use of virtual reality, suffer from rigid facial expression

which is hard to assess from the face of a person using gear that covers large parts of the face. Emotions reliably obtained from the kinematics of other parts of the body could be added procedurally to provide avatars not just with bodily expression but also with facial expression.

Motion style, including the body kinematics that express emotion, is often transmitted by minute changes in pose and dynamics. Hands are likely to play an important role in that context. To study hand motion on the stylistic level, we need better models that describe individual shape, pose and motion of hands.

James Z. Wang: As we have discussed in our recent ARBEE work [6], breakthroughs are needed in both computational/technical fields and psychology.

First, vision-based human pose estimation methods are noisy in terms of jitter errors. Emotion understanding demands substantially higher precision of body landmark locations, compared with typical vision applications.

Second, vision-based methods usually address whole-body poses, which have no missing landmarks, and only produce relative coordinates of the landmarks from the pose instead of the actual coordinates in the physical environment. Real-world emotion understanding requires the modeling of the person and the environment together.

Third, while deep learning methods generate better recognition results, they offer minimal explainability and interpretability. In many emotion understanding applications, these characteristics are more important than accuracy. Breakthroughs in core machine learning and statistical modeling are still needed.

Finally, it can be a technical breakthrough to develop an effective data-driven model that can incorporate the person's personality, gender, age, race, ethnic groups, cultural background, and personal characteristics, as well as the context.

5 Vision of the Future

From the perspectives of your field, how do you view the future of this area of research?

Norman Badler: From the perspective of computer graphics motion generation (animation), parameterization is currently the key to automated human (character) behaviors, but the state-of-the-art is not yet sufficient to replace the intuition, skill, and subtlety of computer animators or direct motion capture in the process. Again, I would argue that this is because the transformation from desire to motion expression lacks important input structures. I do not know of any extant computer animation system that allows one to input a character's personality, mood, social context, response history, personal gestural vocabulary, relationship to any interlocutor, and emotional state and then outputs the "right" behavior. I do think this is eventually possible, but it will require deeper understanding of all these influences on behavior which is going to be empirically

challenging without some decent annotational bases for learning the complexities of such mappings and their interactions. People learn such associations from years of personal experience, so it must be possible, but even people aren't flawless in their judgements.

Nadia Berthouze:

- Real-world applications to support people in the wild
- Ubiquitous sensors sensing the body in its details

Rick O. Gilmore: I am both optimistic and pessimistic. My optimism stems from the considerable energy, creativity, and inventiveness of researchers in AI. My pessimism stems from the fact that most of the computer vision models I've used are not especially valuable to researchers in the behavioral sciences, and there is not enough collaborative work across these communities to share expertise.

I am of both minds also when I think about the future of data. We need bigger, better, more carefully annotated, and more widely shared data about real-world behavior. The Play & Learning Across a Year (PLAY) project (play-project.org) is one effort to collect data analogous to a "human behaviorome." But COVID-19 has put the effort on hold for now.

Kerri L. Johnson: This is truly an exciting time to be conducting research in this area. The very integration of vision, computational, and social science approaches to understanding dynamic movement is happening, albeit slowly. This has the potential to provide insights that have heretofore been impossible due to computational processing limits and data sets of scale.

Agata Lapedriza: Collecting annotated data is complicated and expensive. I think we need to think about self-supervised approaches or cross-modal approaches for emotion recognition.

Xin Lu: I spent most of my recent years developing machine learning algorithms, systems, and a real-world mobile camera application (the Photoshop Camera). I believe bodily expressed emotion is going to be more fascinating in the mobile and Internet of Things (IoT) era, which has already arrived. Everyone spends more and more time on mobile and IoT devices every day, and people will be interested in knowing more of themselves by way of these devices. Mobile or IoT devices can privately capture personal data and identify useful and interesting patterns without sharing these data to the cloud. The wide availability of these personal devices as well as their powerful computational power make up an unprecedented opportunity to study sensitive and private information such as bodily expressed emotions of individuals.

Nikolaus Troje: I have always been fascinated by the divide between "bodily expressed emotion" and cognitive-rationale responses. For instance, my current

work involves immersing experimental participants in virtual reality where we simulate fearful situations. Participants are perfectly aware of the simulated nature of the situation and the fact that they are safe. Nevertheless do they show somatic responses that express true fear: For instance, when navigating a deep abyss on a narrow plank they are sweating, their hard beat increases and their behaviour changes to a point where they are no longer able to function normally [2].

Emotions are to a large degree somatic responses and they express themselves in ways that often are not accessible to cognitive control. Asking someone how she feels is not providing the same information as the one we get when assessing bodily responses. Understanding the latter opens avenues toward treatment of emotional disorders (*e.g.* PTSD, specific phobias). It is also important to assess the affect of fearful situations in the real world, but also in the context of gaming and entertainment, particularly in virtual reality.

A very interesting future area of application is the design of autonomous avatars that also convey emotion in convincing ways.

James Z. Wang: While a comprehensive understanding of human bodily expressed emotion is daunting, I believe we will start to see exciting industrial applications with some limited recognition capabilities. For example, if we limit the emotion categories to a few common emotions, the problem becomes a lot more tractable. As larger and more extensive datasets and more efficient and effective computational modeling techniques become available, the accuracy level and comprehensiveness of the computer-based understanding will continue to improve. When we were developing machine learning and statistical modeling based image annotation in the early 2000s [4], many researchers did not think it was possible to achieve a usable level of accuracy. Today, less than just two decades later, we have already seen wide industrial adoption of machine annotation of images. While emotion understanding is arguably much more subtle and hence more difficult to model than recognizing ordinary objects, I'm optimistic that over time, with increased multidisciplinary collaboration, we will be able to see a number of exciting applications that can improve people's lives.

6 Advice on Innovation

What advice would you offer young researchers trying to work on this topic? How can they be innovative?

Norman Badler: Set yourself a long term goal and evaluate how to proceed toward that goal starting from known techniques combined with integrated insights. For example, don't avoid reading psychology or kinesiology literature. At lot of it won't be useful, but you may get insights from reading outside your academic box. Most of my (ultimately) important computational insights came from works in Natural Language Processing, dance, ergonomics, and communicative and manipulative behaviors. Rather than shun complexity in favor of small

piecemeal steps, see if you can manage complexity through problem decomposition and combination. CG is a poster child for combining techniques to achieve stunning results by mixing image processing, ML, algorithmic techniques, human perceptual constraints, and model representations. Animation is moving slowly toward complexities but is mostly still a technique-based discipline. Computer vision is also likely to appreciate that interconnecting and integrating smaller pieces of a technology puzzle into a psychologically justifiable larger framework may be more fruitful than trying to build a single whiz-bang n-layer network with miracle outputs. That's going to take a longer view but I think it's worth trying.

Nadia Berthouze: By thinking to the applications and purpose of use. Understanding the purpose and the context of use may lead to think differently to how we approach the problem. This happened to us in the context of physical rehabilitation. It led to understand the needs for it, its multiple uses, contexts of use and barriers to its use and to challenge current approaches to the problem.

Rick O. Gilmore: There is a huge but largely untapped pool of knowledge in the community of behavioral scientists that could and should inform research in this area. So, my advice would be to make friends with your colleagues in psychology, ethology, communication sciences, and education research.

Kerri L. Johnson: My advice is to follow your interests, not necessarily what seems trendy at the time. That's where you will be your most innovative. Ask the questions that pique your curiosity, and do whatever is necessary to conduct the most rigorous research that will answer them. Be meticulous and careful in your design, analysis, and reporting. And balance your research portfolio with projects that are "reach" projects (long term, higher risk) and those that are safer and faster.

Agata Lapedriza: I would recommend students that want to work in emotion recognition to read a lot of psychology papers and books. We are trying to create machines that do something that is very challenging even for us: we are not that good in inferring other's emotions, and sometimes we are not even able of categorizing how we feel. Being aware of the findings and theories of emotions from a psychology perspective is very important.

Xin Lu: First, think deeply and read broadly. Emotion recognition/analysis is still at an early stage, and an ideal formulation of this problem doesn't exist to a certain extend. Therefore, it is essential to think deeply on fundamentals. Meanwhile, reading broadly to understand theories and methods by way of well-defined problems builds up tools for young researchers to better solve complex and abstract problems.

Second, learn programming tools. As the volume of data is soaring, having a strong expertise in programming is essential for young researchers to be innovative in the digital and mobile world.

Third, scope the problem. An essential skill that young researchers in this area need to have is scoping the problem to one that is solvable. While the sky is big and tool sets are rich, we cannot throw a tool randomly to the sky and expect that a bird is caught. Rather, we need to navigate the tool that pivots towards the scoped problem.

Nikolaus Troje: I think the most interesting aspect of emotion research is the dynamics of emotion emerging in social situations. If two people are communicating, they may enter the encounter with certain emotions that reflect their history until that point, but the really interesting processes are the ones that happen during that interaction. In order to study such processes we need to understand the sophisticated body language between the two interlocutors, we need to understand how it expresses the emotional state of the acting person, and how it affects the emotional state of the perceiving person.

Emotion is nothing static, it is reaction to the (social) environment, on the one hand, but also affects that environment, on the other hand. In order to study the emotional progression of a dyadic encounter we need to be able to monitor both interlocutors simultaneously, and we need to come up with dynamic models that can describe that progression. Of course, things would become even more interesting, if we were able to study emotional progression in larger groups.

In the previous section, I talk about communication with autonomously behaving avatars. Again, for such tools to become truly convincing we have to find ways to not just model emotion in itself, but we have to understand the interactive emotional dynamics emerging during social interactions.

James Z. Wang: First, I'd encourage young researchers interested in this topic to collaborate with those with complementary expertise and take a multidisciplinary approach. The BEEU workshop aims to bring active researchers from different fields together. I believe we are only seeing the beginning of a lot of fruitful collaborations between computer science and behavior sciences.

Second, because this problem is not a typical computer vision or machine learning problem, and it is complex and not well-defined, young researchers need to have the courage to face significant challenges and even many failures. It is likely not going to be a project that can quickly generate publications or citations. It needs patience and perseverance to accomplish novel and significant results that have lasting impacts.

7 Acknowledgments

The opinions expressed in this article are panelists' personal views. J. Z. Wang is supported by the National Science Foundation under Grant No. 1921783 and

the Amazon Research Awards Program. Reginald B. Adams, Jr. and Yelin Kim contributed to the discussions related to the theme of this panel.

References

1. Byun, M., Badler, N.I.: Facemote: qualitative parametric modifiers for facial animations. In: Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation. pp. 65–71 (2002)
2. Eftekharifar, S., Thaler, A., Troje, N.F.: Contribution of motion parallax and stereopsis to the sense of presence in virtual reality. Journal of Perceptual Imaging (2019)
3. Ghorbani, S., Mahdaviani, K., Thaler, A., Kording, K., Cook, D.J., Blohm, G., Troje, N.F.: MoVi: A large multipurpose motion and video dataset. arXiv preprint arXiv:2003.01888 (2020)
4. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(9), 1075–1088 (2003)
5. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 1–16 (2015)
6. Luo, Y., Ye, J., Adams, R.B., Li, J., Newman, M.G., Wang, J.Z.: ARBEE: Towards automated recognition of bodily expression of emotion in the wild. International Journal of Computer Vision **128**(1), 1–25 (2020)
7. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5442–5451 (2019)
8. Zentner, M., Grandjean, D., Scherer, K.R.: Emotions evoked by the sound of music: characterization, classification, and measurement. Emotion **8**(4), 494 (2008)