On Recovering Structure of Affect

Ashish Kapoor[1], Mary Czerwinski[1], Diana Lynn MacLean[2] and Alex Zolotovitski[3]
[1]Microsoft Research Redmond, [2]Stanford University, [3]Medio Systems Inc.
{akapoor, marycz}@microsoft.com, malcdi@stanford.edu, alex.zolot@gmail.com

***Abstract*—This paper presents novel human computation experiments geared towards uncovering the structure of affect. Using Mechanical Turk workers across 2 separate studies, we empirically verified some of the popular beliefs about the structure of affect, but also provide some new evidence. We replicate and reveal not only the statistical structure of the dimensions of affect, but also the effect of cultural influences. We close with a proposition for a framework for doing this kind of large scale research and provide recommendations and opportunities for innovations in research around emotional theory.**

Keywords—Affect, Models of Affect, Dimensionality Reduction, Clustering, Mechanical Turk

I. INTRODUCTION

The literature on emotion is rife with strong opinions around emotional theory. Theories on emotions can be traced back to the ancient Greeks, such as Plato and Aristotle. Today, there are theories spanning psychology, neurology, somatic theories, situated perspective and evolutionary theories, etc. However, researchers doing affective computing have tended to focus on a small number of theories or classifications in order to build systems that model emotional behavior. The most common models include Russell's circumplex model [1], Ekman and Freisen's [2] model based on discrete sets of universal emotions and Plutchnik's [3] emotion wheel. These three models provide useful concepts that researchers in affective computing have used for modeling purposes, whether they are entirely accurate or not. In particular, the circumplex model has been heavily leveraged in the modeling of emotion (see [4] and [5] for reviews).

As the literature on the theoretical explanations is so vast and full of debate, and given that these models were proposed several decades ago, we were motivated to approach this topic from a different angle. The key motivation behind this work is the fact that there are several methods and tools available to researchers that resulted from significant advances in computing and data analysis in the last few years. We enjoy data collection frameworks that enable us to collect data at a scale that was not possible earlier: In particular, methods like crowdsourcing and human computation enable us to recruit large numbers of human participants, and we seek to explore how such big-data collection capabilities can help us validate older theories on emotion and discover previously unknown aspects of such models.

Further, there also has been a significant advancement in the field of machine learning and statistical data understanding, and this promises deeper understanding of the nature of affect. In particular, recent advances in clustering and manifold learning enable us to do a much more thorough empirical analysis of the problem. Such tools, when combined with the possibility of collecting a large amount of data via crowdsourcing, provide us with a unique perspective on the problem that was not possible earlier. There are three core contributions of this paper that highlight the potential of combining crowdsourcing, human computation and modern machine learning:

1. The paper demonstrates a novel human computation experiment that is geared towards uncovering the structure of affect.
2. We use this large scale data collection not only to empirically verify some of the popular beliefs about the structure of affect, but also to discover previously unknown aspects. Specifically, we discover (a) correlational structure between aspects of affect, (b) the inherent dimensionality of the space and (c) changes in the structure of the space with context.
3. Finally, we propose a framework that enables such large scale research and conclude with sets of recommendations and possible opportunities for exploring promising new research directions.

II. RELATED WORK

Crowdsourcing and human computation have had deep impact on several fields. Crowdsourcing is a way to recruit large numbers of humans in order to get a task completed, using an open call for work with small amounts of contractual requirements. Well known examples include large projects like Linux, Wikipedia, etc. Of particular interest is the crowdsourcing work done by [6] on Emo20Q, wherein the crowd was used to gather textual questions related to emotional terms. This work identified classes of questions relating to emotions that could be used to drive a system that would respond in an emotionally appropriate manner.

Human computation, on the other hand, uses the crowd to perform micro task units using monetary incentives to recruit users. Human computation can include games like Fold-it, TagATune and ESP [7–9] to solve large, often intractable problems in computer science or biology, for example. While these games are self-motivating, other systems like Amazon's Mechanical Turk (MTurk; www. https://www.mturk.com/mturk/welcome) utilize monetary incentives to entice workers. These workers can also be motivated by improving their reputation on certain kinds of

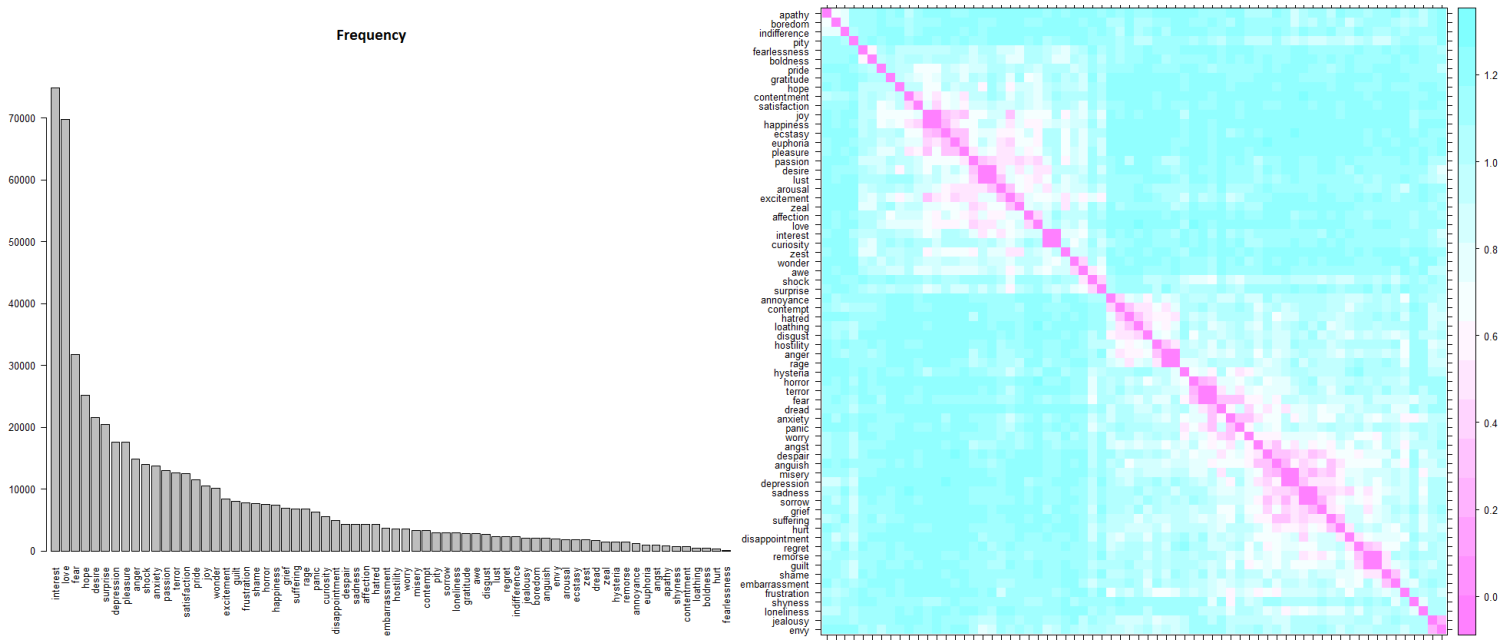

Figure 1. (Left) The emotion terms being used in this study and their associated usage frequencies as per the Corpus of Contemporary American English. (Right) The distance matrix between these terms recovered via the Mechanical Turk studies.

tasks with respect to their peers in order to get more/more profitable and interesting work. We decided to use MTurk to study how individual users in the United States think about the psychological similarity of the emotional terminology space.

It's worth noting that early work on the empirical modeling of affect focused on first obtaining data from a small number of humans and then running computational analysis. For example, several studies focused on similarity-dissimilarity of facial/verbal expressions, and performing analysis using multidimensional scaling. These studies resulted in very similar two-dimensional structures being reported across researchers [10–12]. It was found that the cognitive representation of affect is best described by the pleasantness-unpleasantness and arousal-sleep dimensions, accounting for most of the variance in the judged similarities [13]. Russell's classic [1] paper used similar studies to replicate this two or three dimensional representation using a specific set of 28 affective terms and concluded that the affective words were seen as some combination of the pleasantness/arousal axes, much as Schlosberg's original idea of a circular order of affective terms in a 2D space[10].

III. Theories of Emotion

Many theories of emotion have considered each affective concept as a separate dimension as espoused originally by Nowlis (e.g., [14]) and later by others well known in the field, such as Izard's [15] theory of discrete emotions and Ekman's (e.g., [16]) cross-cultural work on the facial expression of emotion. This basic theory of emotion movement also led to self-report instruments which were commonly used to assess affect in psychology (e.g., [15]), some (and their antecedents) of which are still in use today.

One of the most widely used models for studying emotion is Russell's Circumplex Model of Affect [1]. The model was put forward to debunk prevailing psychological theories of affect that characterized emotions as discrete, limited in number, independent of each other, and driven by separate neurophysiological systems. According to this model, differences and similarities between affective states are modeled via two orthogonal and bipolar dimensions—valence on the horizontal axis (pleasure/displeasure) and activation or arousal (low/high) on the vertical axis. Different affective states are considered to be blends of these two dimensions. For example, excitement and enthusiasm would be considered combinations of pleasure and high arousal, while boredom and depression would be a mixture of displeasure and low arousal.

The psychological and neuropsychological communities have long relied on the basic theories of emotions approach to guide their thinking, but there is a recent trend to move toward Russell's circumplex model in those communities as well [5], [17], [18] though there have been some limitations to its use noted, e.g., [17], [19]. As argued by Posner et al, the circumplex model has many strengths, in that it is 2 (or 3) - dimensional, well known, well understood, and can account for many of the newer findings in neuropsychology and psychiatry [4]. In addition, efforts have been made to design computer tools to measure affect using the model (e.g., [20]) and others have used it to predict affective choice outcomes in consumer settings (e.g., [18], [21]). This is important to point out, as the model provides researchers with methods for looking at behavioral decision making and potentially, long-term behavioral change.

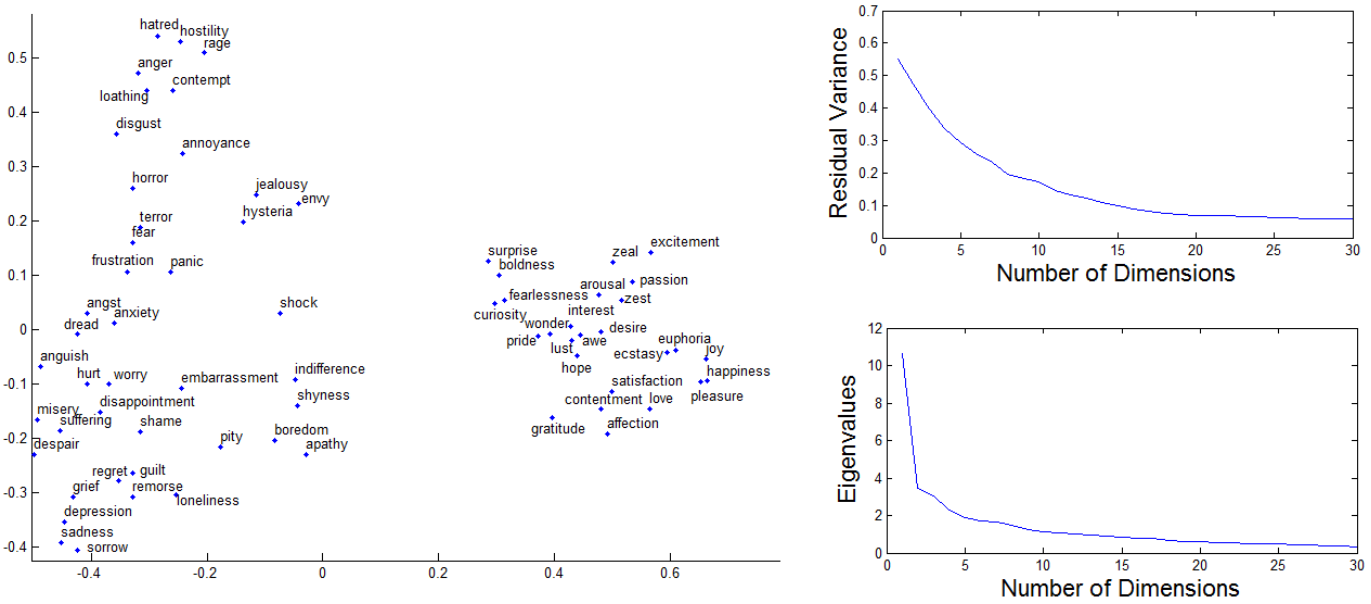

Figure 2. Left figure shows the projection of the emotion words using the top two dimensions found by Kernel PCA. Right figure shows residual variance and eigenvalues associated with the rest of the discovered dimensions.

IV. The Core Framework

Our aim in this work is to explore and understand the nature of affect space. Consequently, we chose 68 different words that are most frequently used to describe an affective state. These terms were chosen by mining the web and considering their frequency usage. Figure 1 (left) shows 68 of these words and their frequency of usage as documented in Corpus of Contemporary American English. We conducted all our studies on these 68 emotion terms.

There are two key components of the framework: first, is the potential of collecting large amounts of data via Mechanical Turk (Mturk). Mturk enables surveying and recording responses from human participants from all over the world at a fraction of a cost previously. In comparison, such experiments in lab settings are not only tedious and expensive, but also very hard to scale to a large population. Further, one big advantage of our framework is that collected data can be easily analyzed by variables such as age/gender/geographic location, that are known to significantly affect the structure of affect. Consequently, it is now possible to study such differentials across cultures/genders/age, etc., and test the hypothesis that the structure of affect is universally constant.

One of the challenges in running such studies, either Mturk or otherwise, is the fact that the questions being asked should be unambiguous and should lead to consistent answers when asked multiple times to the same participants. To this end, we use a very simple strategy in this work where we ask the participants questions about a pair of emotions, and ask them to compare them in terms of similarity. Note that this task by itself is fairly unambiguous and simple and also participants are likely to answer these questions consistently. Further, this study will easily run on Mturk, while it is extremely hard to run in a traditional setting due to the fact that the large number of total pairs of such tasks is huge (2346 in total).

Once the data is collected, we process the data to compute a similarity matrix (or a kernel) between each of these individual terms. This kernel matrix captures the similarity between all possible pairs and in essence has a lot of information about the structure of the affect space. In particular, the notion of similarity or the kernel matrix is very useful for analysis using recent Machine Learning techniques. We specifically utilize Kernel Principal Component Analysis (KPCA) [22] which is designed to recover latent dimensions of the metric space from similarity matrices. Also, note that the similarity matrix allows us to compute a distance metric (denoted as $d(i,j)$ between any two pairs of affect terms: $d(i,j) = k(i,i) + k(j,j) - 2k(i,j)$. This distance then can be used in any clustering algorithm to discover natural groupings of emotion words. Below we describe two user studies that use these concepts to analyze the space of affect.

V. User Study 1: Structure of Affect

A. Methodology

Using the above described list of emotional terms, we paired all pairwise comparisons of each emotional term with each other (including the identical word pairings as a check for good performance by the workers). Starting from a base of 68 emotional terms, we ended up with 2346 pairs of terms that we collected similarity ratings for, recruiting 10 participants per rating. We used a 5-point Likert scale, with 1=very dissimilar, and 5=very similar. The identical word pairings were included as a foil—if the participant didn't rate the identical word pairs as a 5, their data could be thrown out. We only received one rater that didn't rate an identical word pair as a 5, so we kept all of the data. In addition to asking for a similarity rating on each word pair, we also asked participants to tell us which of the two words was more intense (higher arousal) and which word was considered to be more positive.

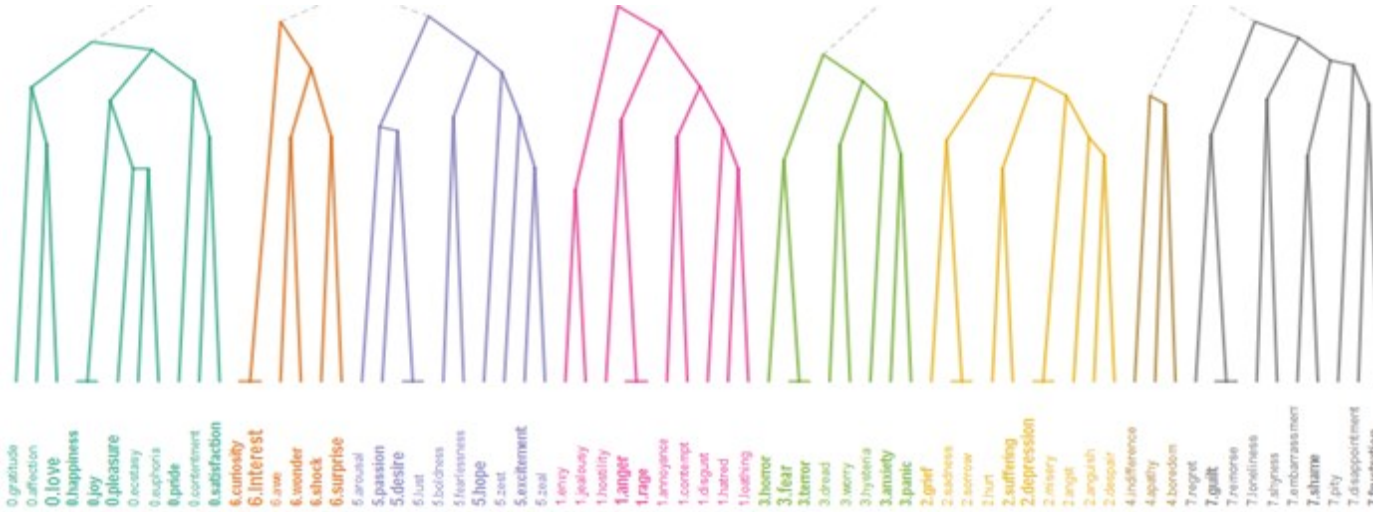

Figure 3. Cluster dendogram based in distance metric recovered via the Mturk Study. The data indicates that there the emotion words can be clustered into eight different groups based on the response of the participants.

B. Participants and Apparatus

Participants were required to be from the United States only, have a 95% approved reputation or higher, and had to have completed at least 50 MTurk tasks before. Participants were paid 6 cents a task and ratings took 16 seconds, on average, for an hourly rate of over 6.00/hour USD, on average. It took 48 hours for 10 ratings per term to be complete the study.

C. Analysis and Results

The data collected from the MTurk experiment was used to estimate the kernel matrix. In particular, for every pair of emotion words, the similarity ratings were simply aggregated across all the users to yield the un-normalized kernel with entries $\tilde{k}(i,j)$. This matrix is normalized via the following transformation: $k(i,i) = \frac{\tilde{k}(i,j)}{\sqrt{\tilde{k}(i,i)\tilde{k}(j,j)}}$. This leads to a kernel matrix where the diagonal is one (highest similarity), and rest of the off-diagonal terms are between zero and one.

KPCA was then applied to this kernel matrix and we show the projection of all the 68 emotion words using the top two recovered dimensions. Figure 2 Left shows this projection and we observe that indeed the top two dimensions capture the notion of valence and arousal, as proposed in the literature. However, the surprising thing that we observed is the fact that the variation of arousal when conditioned on valence is not the same. In particular, we see that the range of variation in arousal for negative affect is much larger than for the positive terms. This strongly suggests that the arousal and valence are not independent. In order to test the hypothesis that valence and arousal axis are dependent, we first partitioned the dataset by the sign of the first dimension (i.e. valence > 0 and <= 0). F-test on this partitioned data showed significant differences in the variance along the second dimension (arousal) axis at 95% confidence. Most of the models do not account for such relationships, and perhaps the 2D representation of affect and arousal axis is incomplete. Specifically, in such representations there are regions that are infeasible due to structural constraints between arousal and valence.

Next, one of the most important questions in affect structure discovery is about the number of inherent dimensions. We attempt to answer this question via KPCA as well. In particular, we look at the increasing number of dimensions, and compare how well the representation explains the data. Specifically, we look at two different metrics: residual variance and eigenvalue associated with the individual dimensions. The first quantity measures how well the dimensions preserve the original similarity space, and the second quantity inherently captures the noise to signal ratio of each individual dimension. The plots for both these measures, as dimensions are increased, are shown in figure 2. While a lot of information is contained in the first two dimension it, is clear from these plots that there is strong signal up to the fifth dimension. Below is the partial sorted list (low to high) according to the position on the axis (formatted as First 3, Middle 3 and Last 3):

- Dimension 3: Panic, Terror, Fear, .., Suffering, Pleasure, Desire, .., Boredom, Contempt, Indifference.
- Dimension 4: Lust, Desire, Envy, .., Excitement, Awe, Suffering, .., Boredom, Contentment, Apathy.
- Dimension 5: Curiosity, Surprise, Wonder, .., Fear, Loneliness, Terror, .., Affection, Ecstasy, Love.

Understanding the taxonomy and role of these dimensions is an important area of future work and deserves replication and a more thorough exploration.

Finally, we also explore the possibility of a discrete affect model space via clustering methods. In particular, we induce the distance metric via the kernel as described earlier. The resulting distance metric is shown in Figure 1. We have re-ordered the rows and columns to highlight the natural clustering according to the observed data. We ran a hierarchical clustering algorithm and recovered the dendogram as shown in Figure 3. We observed that the emotions do have a tendency to cluster together; however, one of the challenges of such discrete representation is choosing the correct number of clusters. While it appears there could be eight clusters in our data, relegating those clusters to be useful in both theory and practice might be too complicated. One of the most useful aspects of the circumplex model is its simplicity.

Figure 4. Arousal / Valence plots of (Left) Indian and (Right) US MTurk participants on responses to self-perceived affect (orange) and affect perceived by others (purple). Valence and arousal are represented on X and Y axes respectively. Lines connect dots relating to the same emotion word.

VI. User Study 2: Understanding Differences in Perception of Affect

The second user study aims to explore how people's perceptions of affect change with different frames of reference. Such change in affect perception from one context to another is an interesting consideration that has not been studied much, and may reveal insights about respondents that would otherwise stay undiscovered. Understanding such perception differences of affect due to change in frames of reference could also be an invaluable diagnostic tool. For example, depression (especially mild depression) is extremely difficult to diagnose. A depressed person's responses on an affective questionnaire might fall within normal population parameters; if we compared their responses regarding their own affective state, however, to their impressions of others' affective states, and there is a significant difference, then it might warrant further investigation.

A. Methodology

We conducted a study on Amazon's Mechanical Turk in which we asked Turkers to rate the arousal and valence of the same 68 emotion words as in Study 1. In this study, we considered four conditions, determined by participants' geographic location and by whether they were asked to rate each word in terms of how it made them feel, or in terms of how it made others feel (see Table 1). Each word was rated 10 times by different workers in each condition (40 times in total). Words were rated on a 7-point Likert scale (1=extremely positive, 7=extremely negative for valence; 1=extremely active, 7=extremely passive for arousal).

B. Participants and Apparatus

Participants were required to be from the United States or India only. And, as in Study 1, the participants were required to have a 95% approved reputation or higher, and had to have completed at least 50 MTurk tasks in the past.

Table 1. The Four Conditions for User Study 2

	Self	Other
India	When you experience anger, how positive or negative do you feel? What would be your level of arousal (low to high)?	When someone around you experiences anger, how positive or negative do you think they feel? What would be their level of arousal (low to high)?
United States		

C. Analysis and Results

Figures 1 and 2 show "affective spread" for the self and other conditions in the Indian and US worker population, respectively. At first glance, it's interesting to see that regardless of whether you are rating an emotional term for yourself or for others, the two countries rate the emotional words in very different quadrants of the 2x2 circumplex model. The US data falls into the upper left, lower left and upper right quadrants almost exclusively, while the Indian data falls primarily within the upper right or lower left. This could indicate that the Indian culture does not experience highly aroused, negative affect, nor possibly very low arousal, positive feelings. The US data similarly reflects a lack of ratings in the low arousal, positive space. More importantly, the US data shows more extreme values of ratings from "self" to "other". The Indian data does not seem to change too dramatically. This is shown in Figures 4. Though there were several emotion words ranked similarly in each plot, there were also notable differences. To explore this further, we determined emotional terms for which participants rated as pertaining to themselves much different from pertaining to others. Listed below are the 5 most different self/other ratings by US workers:

1. **Horror:** Others are highly aroused (when they experience horror); my arousal levels are neutral.
2. **Disappointment**: Others have very low arousal (when they experience disappointment); I am neutral.
3. **Embarrassment**: Others appear to be slightly un-aroused (when they are embarrassed); I am neutral.
4. **Fear**: Others seem less aroused and less negative (when they experience fear); I am slightly aroused and only somewhat negative.
5. **Sorrow**: Others appear to be less aroused and less negative when they experience sorrow; I am only mildly un-aroused.

For comparison, here are the 5 most different self/other ratings by Indian workers. In all cases but the last (shock), others' arousal is rated as less than one's own.

1. **Loathing:** Others seem more positive and more aroused when they experience loathing.
2. **Pity:** Others seem more aroused when they experience pity and less positive.
3. **Embarrassment:** Others seem more positive and more aroused when they are embarrassed.
4. **Contempt:** Others seem to be more positive and more aroused when they experience contempt.
5. **Shock:** Others seem less positive and less aroused when they are shocked.

It would appear that there is the greatest discrepancy between self/other ratings when the emotion words are associated with negative valence. Both countries show a large discrepancy between self and other ratings for *embarrassment*, for example. Interestingly, the US ratings showed more extreme differences from self to other ratings than did the Indian ratings. One possibility for explaining this is that it is easy to know when you are experiencing a strong, negative emotion, but to recognize it in someone else, the emotional expressions they would have to convey must be highly salient. The US ratings could in some cases be a reflection of that. There may be other, cultural issues driving the ratings differences between the two countries, and for the first time we can actually explore them and begin to form hypotheses.

VII. Discussion and Future Work

A new era of research is upon us when we can literally design and run a study on emotion in the matter of 48 hours, gaining access to much more data, and hence, power, from a statistics point of view, than ever before. Of course, it is of the utmost importance that good screening techniques and "test" questions are included in the jobs submitted to Mechanical Turk in order to ensure quality responses, but it has been our experience that most Turkers enjoy doing these types of studies and want to keep their reputation high. We had to throw out less than 1% of our data for these studies, which is remarkable. Gaining access to workers from different cultures has also been an extremely valuable asset and because of this we have been able to expose new findings around the psychological space of emotions.

Our experiments also indicate that while arousal and valence are indeed two key dimensions, they are not necessarily independent, and that current representations are incomplete. Further, our data indicate evidence of the existence of more than two key dimensions. There is also a feasibility of exploring discrete representations of affect via clustering that needs to be explored further. Finally, we have shown that the structure of affect might change based on one's perspective, cultural or otherwise. Our work points to the possibility of discovering richer and better models of affect by utilizing recent data acquisition and analysis tools.

References

[1] J. A. Russell, "A circumplex model of affect.," *Journal of Personality and Social Psychology*, vol. 39, no. 6, 1980.

[2] P. Ekman and W. V Friesen, "Constants across cultures in the face and emotion.," *Journal of Personality and Social Psychology*, vol. 17(2), 1971.

[3] R. Plutchik, "A general psychoevolutionary theory of emotion," *Emotion: Theory, research, and experience*, vol. 1, no. 3, 1980.

[4] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology.," *Development and Psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[5] N. A. Remington, L. R. Fabrigar, and P. S. Visser, "Reexamining the circumplex model of affect," *Journal of personality and social psychology*, vol. 79, no. 2, pp. 286–300, 2000.

[6] A. Kazemzadeh, S. Lee, P. G. Georgiou, and S. S. Narayanan, "Emotion Twenty Questions: Toward a Crowd-Sourced Theory of Emotions," *Affective Computing and Intelligent Interaction*, vol. 6975, pp. 1–10, 2011.

[7] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and F. Players, "Predicting protein structures with a multiplayer online game.," *Nature*, vol. 466 (7307), 2010.

[8] E. Law and L. Von Ahn, "Input-Agreement : A New Mechanism for Collecting Data Using Human Computation Games," *Machine Learning*, vol. pp, pp. 1197–1206, 2009.

[9] L. Von Ahn, *Games with a purpose*, vol. 39, no. 6. IEEE Computer Society Press, 2006, pp. 92–94.

[10] H. Schlosberg, "The description of facial expressions in terms of two dimensions.," *Journal of Experimental Psychology*, vol. 44, no. 4, 1952.

[11] R. P. Abelson and V. Sermat, "Multidimensional scaling of facial expressions.," *Journal of Experimental Psychology*, vol. 63(6) 1962.

[12] N. Cliff and F. W. Young, "On the relation between undimensional judgments and multidimensional scaling," *Organizational Behavior and Human Performance*, vol. 3, no. 3, pp. 269–285, 1968.

[13] J. A. Russell, "Evidence of convergent validity on the dimensions of affect.," *Journal of Personality and Social Psychology*, 36(1) 1978.

[14] V. Nowlis and H. H. Nowlis, "The description and analysis of mood," *Annals of the New York Academy of Sciences*, 65(4)1956.

[15] C. E. Izard, *Human emotions*. Springer, 1977.

[16] P. Ekman, "Universals and cultural differences in facial expressions of emotion.," in *Nebraska symposium on motivation*, 1971.

[17] P. Ekkekakis, "Affect circumplex redux: the discussion on its utility as a measurement framework in exercise psychology continues," *International Review of Sport and Exercise Psychology*, 2008.

[18] G. Suri, G. Sheppes and J. J. Gross, "Predicting Affective Choice," 2012.

[19] D. C. Rubin and J. M. Talarico, "A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words," *Memory*, vol. 17(8), 2009.

[20] I. A. Khan, W.-P. Brinkman, and R. M. Hierons, "Towards a Computer Interaction-Based Mood Measurement Instrument," *Proc. PPIG2008, ISBN*, pp. 971–978, 2008.

[21] D. McDuff, R. el Kaliouby, D. Demirdjian, and R. Picard, "Predicting Online Media Effectiveness Based on Smile Responses Gathered Over the Internet," *IEEE Conference on Automatic Face & Gesture Recognition*, 2013.

[22] S. Mika, B. Schölkopf, A. Smola,K. R. Müller, M. Scholz & G. Rätsch. "Kernel PCA and de-noising in feature spaces", in *Advances in neural information processing systems* 1999.

[23] C. Kaernbach. "On dimensions in emotion psychology", IEEE Conference on Automatic Face & Gesture Recognition, 2011.