Investigating the Evaluative Dimensions of a Large Set of Communicative Facial Expressions: A Comparison of Lab-Based and Crowd-Sourced Data Collection

Dilara Derya, Korea University, Republic Of Korea
Ahyoung Shin, Korea University, Republic Of Korea
Haenah Lee, Korea University, Republic Of Korea
Christian Wallraven Korea University, Republic Of Korea

Asian Conference on Psychology and Behavioral Sciences 2015
Official Conference Proceedings

Abstract
Facial expressions form one of the most important non-verbal communication channels. Although humans are capable of producing a wide range of facial expressions, research in psychology has almost exclusively focused on the so-called basic, emotional expressions (anger, disgust, fear, happy, sad, and surprise). Research into the full range of communicative expressions, however, may be prohibitive due to the large number of stimuli required for testing. Here, we conducted both a lab-based and an online, crowd-sourcing study in which participants rated videos of communicative facial expressions according to 13 evaluative dimensions (arousal, audience, dominance, dynamics, empathy, familiarity, masculinity, naturalness, persuasiveness, politeness, predictability, sincerity, and valence). Twenty-seven different facial expressions displayed by 6 actors were selected from the KU Facial-Expression-Database (Shin et al., 2012) as stimuli. For the lab-based experiment, 20 participants rated all 162 (randomized) video stimuli. The crowd-sourced experiment was run on Amazon Mechanical-Turk with 423 participants, selected as to gather a total of 20 ratings per stimulus. Within-group reliability was high for both groups (r_Lab=.772, r_Mturk=.727 averaged across 13 dimensions), with valence, arousal, politeness, and dynamics being highly reliable measures (r>.8), whereas masculinity, predictability, and naturalness where comparatively less reliable (.3<r<.6). Importantly, across-group correlations showed a highly similar pattern. Our results first show that it is feasible to conduct large-scale rating experiments using crowd-sourcing stimuli. Additionally, the ratings paint a complex picture of how facial expressions are evaluated. Future studies will use dimensionality analyses to further investigate the full space of human communicative expressions.

Keywords: Crowdsourcing, Facial expression recognition

iafor
The International Academic Forum
www.iafor.org

Introduction

Communication is a crucial element for building and sustaining a functional society. Communication in humans can be either verbal (that is, using the contents of spoken language), nonverbal (that is, using no-speech signals such as body language, prosody, and facial expressions), or both. Among the non-verbal signals, facial expressions are one of the most important ways of communication for humans.

The importance of facial expressions was already highlighted by Charles Darwin, when he suggested that facial expressions can be hard to suppress when certain muscles that take effort to activate are engaged. Darwin also suggested that facial expressions are used to communicate basic biological needs. During the 1960s, Paul Ekman conducted seminal research that investigated whether, indeed, facial expressions may be shared across cultures. Ekman found and described six, so-called "universal" or "basic" facial expressions that are equally well recognized across different cultural backgrounds (Ekman et al. 1969). These expressions are anger, disgust, fear, happy, sad, and surprise.

Although humans use a much wider range of emotional and non-emotional facial expressions, studies in psychology and related fields have almost exclusively focused on these six facial expressions. Only recently have researchers started to investigate this broader range of facial expressions and have been able to demonstrate reliable recognition of not only emotional, but also conversational facial expressions (Kaulard et al. 2012).

In addition, most studies on facial expressions have focused on analyzing static data, that is, pictures of the expressions. However, several studies comparing static and dynamic stimuli suggest that perception of facial expressions with dynamic stimuli not only yields different performance, but also changes the pattern of results (Cohen et al. 2003; Sandbach et al. 2012; Biele and Grabowska 2006; Kamachi et al. 2001; O'Toole et al. 2002). This difference in performance also has been shown to extend to conversational facial expressions when comparing static and dynamic processing (Cunningham and Wallraven 2009).

One of the reasons that has hampered progress in investigating the full range of facial expressions is the sheer amount of resources that are required for testing: existing databases (Kaulard et al., 2012), for example, contain over 50 expressions from 20 actors – the number of stimuli that need to be tested, therefore, goes into the thousands and represents a big challenge in traditional, lab-based experimental settings. One of the solutions for this problem has come with the advent of crowd-sourcing or online experiments, which promises to allow for testing of large number of stimuli in large populations (e.g., Morris et al. (2011)). In order to be able to use crowd-sourcing, however, it firsts need to be proven that it may produce comparable results to lab-based settings – especially given the more uncontrolled conditions that online experiments contain.

So far, a few studies have recently evaluated the accuracy of crowd-sourcing and found reliable results. Casler et al. (2013) compared a lab-based and a crowd-sourced setting for recognition of novel and familiar objects and found that results were indistinguishable between both groups (even though crowd-sourced participants were somewhat more diverse). Saunders et al. (2013) compared lab experimentation and crowd-sourcing for the purpose of annotating video clips and found reliable results as well. Testing on a more cognitive level, Holden et al. (2013) investigated the test-retest reliability of a personality test and found similar values to lab-based settings. In the context of facial expressions, McDuff et al. (2012), for example, analyzed facial responses to online videos, being able to analyze over 3,000 face videos over 2 months, demonstrating reliable relationships between head movement and facial behavior among other things. Mahmoud et al. (2012) used a crowd-sourcing approach to recreate Darwin's emotion experiment, being able to re-produce previous results about ratings of different kinds of emotional pictures from Darwin's manuscript.

Given the encouraging results of these previous studies, here, we report a validation experiment in which we compare a lab-based and online crowd-sourced facial expression recognition experiment. The experiment uses 27 different facial expressions displayed by six actors, and participants were asked to rate each expression according to 13 evaluative dimensions. Importantly, the experiment uses dynamic stimuli of both emotional and conversational facial expressions to ensure ecological validity. Our goal was to test how well the lab-based results (which already approached the limits of what is doable in a standard testing environment) would match the crowd-sourced results.

Methods

To keep the experiments at a reasonable duration, we selected 27 facial expressions from the KU Facial Expression Database (Shin et al., 2012), spanning a wide variety of communicative and emotional signals. Each expression was available as a video sequence and was performed by 6 different, native-Korean actors (3 male, 3 female). The expressions were elicited by a method acting protocol and validated in several experiments (Shin et al., 2012) – the individual expressions were: showing a considered agreement, "aha" moment (when one suddenly understands something), anger, arrogant (looking down on somebody), being bothered by something, showing contempt towards someone, showing that one is not interested in something ("I don't care!"), disagreeing with something, being disgusted, being embarrassed, reacting in an evasive manner, feeling fearful (terrified of something), a genuine happy laugh, a satiated smile (as if after a good meal), imagining something negative, being impressed by something, feeling insecure, feeling compassion towards someone, feeling pain, being irritated ("rolling your eyes"), remembering something neutral, being sad, showing various kinds of smile (a flirtatious smile, a reluctant smile, a sardonic smile, and a sad/nostalgic smile), being tired. Figure 1 shows peak-frame examples of three expressions.

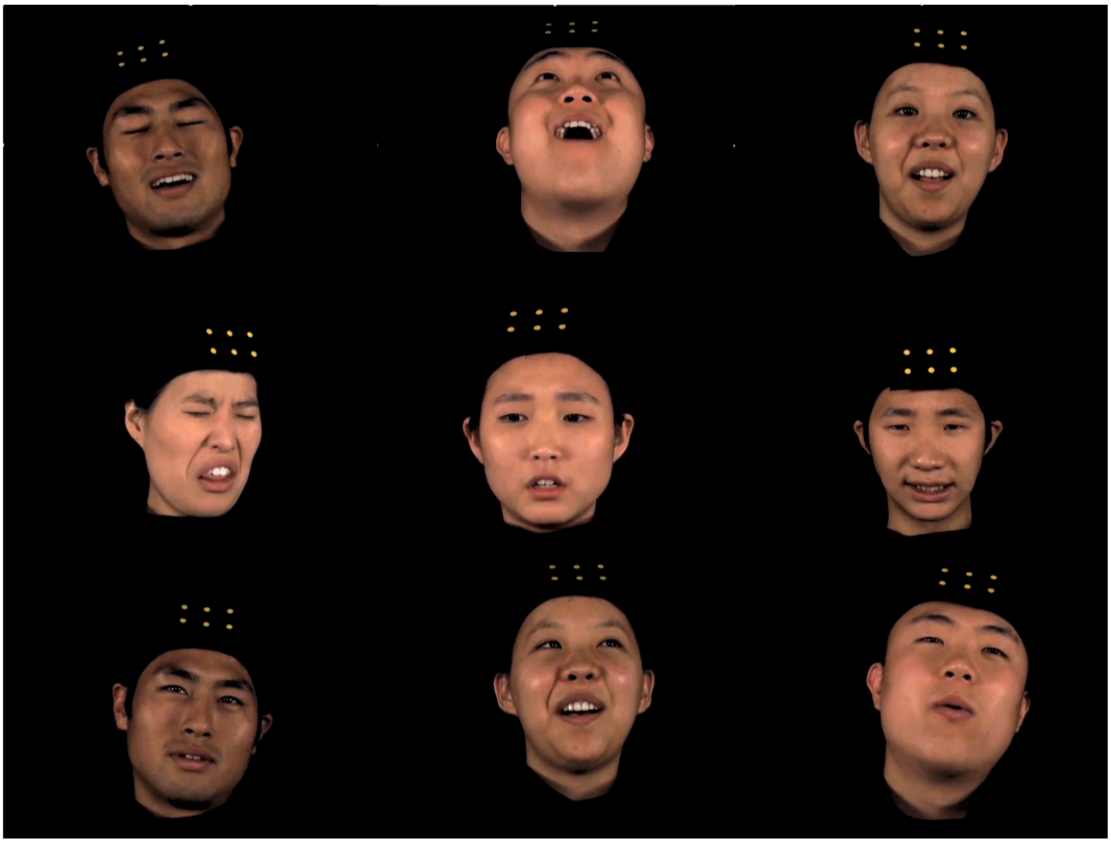
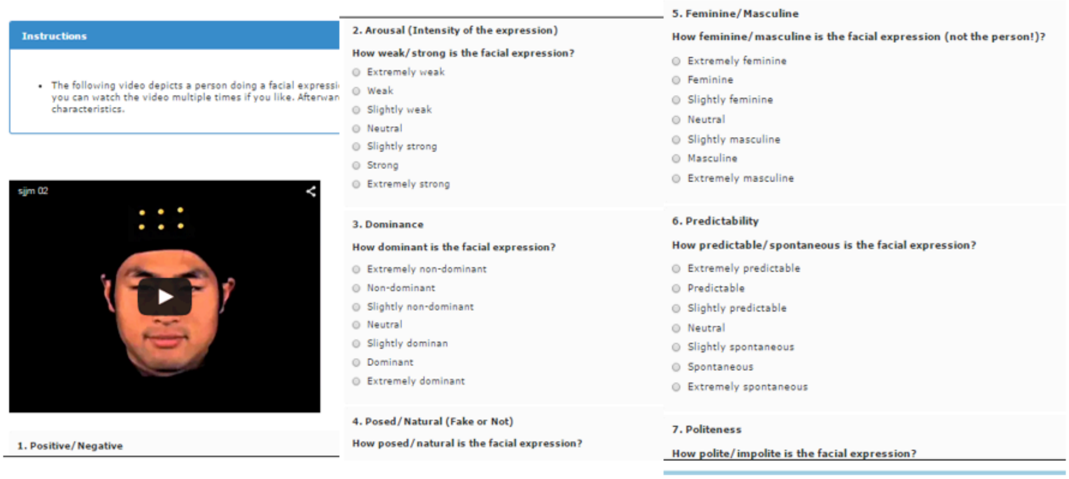


Figure 1: Peak-frames for three expressions from the KU-Facial-Expression Database used in this experiment. First row: “aha” moment, second row: contempt; third row: being impressed by something.

The Amazon Mechanical-Turk website was used to recruit a total of 423 participants, selected as to gather a total of 20 ratings per stimulus. Each participant was allowed to rate a maximum of 10 video sequences. Each video sequence was reimbursed at the rate of .10US$. The lab-based experiment was conducted with 20 participants from Korea University who rated all 162 (randomized) stimuli and participants were reimbursed at around 8US$ per hour.

In each trial, participants had to rate each expression according to 13 different dimensions: arousal = intensity of the expression, audience = needs a conversational partner, dominance = dominates the observer, dynamics = contains a lot of motion, empathy = makes the observer feel empathic, familiarity = is a typical expression, masculinity, naturalness = is a posed or natural expression, persuasiveness = can persuade the observer, politeness, predictability = results from a predictable situation, sincerity, and valence = positive or negative. These rating dimensions were selected based on prior experiments about ratings of emotional concepts (Fontaine et al., 2007; Castillo et al., 2014). Ratings were done on a 7-point Likert-type scale.

The instructions for the crowd-sourcing experiment were explained on top of the webpage (see Figure 2) for each trial. Each participant received a random trial selected from a list. The webpage contained the instructions, the video, and the questionnaire with the rating dimensions. The lab-based experiment proceeded in a similar fashion, except that each participant rated the full list of videos – accordingly, the full experiment here took around 2-3 hours.



Instructions

- The following video depicts a person doing a facial expressi
you can watch the video multiple times if you like. Afterwan
characteristics.

1. Positive/Negative

2. Arousal (Intensity of the expression)

How weak/strong is the facial expression?

○ Extremely weak
○ Weak
○ Slightly weak
○ Neutral
○ Slightly strong
○ Strong
○ Extremely strong

3. Dominance

How dominant is the facial expression?

○ Extremely non-dominant
○ Non-dominant
○ Slightly non-dominant
○ Neutral
○ Slightly dominan
○ Dominant
○ Extremely dominant

4. Posed/Natural (Fake or Not)

How posed/natural is the facial expression?

5. Feminine/Masculine

How feminine/masculine is the facial expression (not the person!)?

○ Extremely feminine
○ Feminine
○ Slightly feminine
○ Neutral
○ Slightly masculine
○ Masculine
○ Extremely masculine

6. Predictability

How predictable/spontaneous is the facial expression?

○ Extremely predictable
○ Predictable
○ Slightly predictable
○ Neutral
○ Slightly spontaneous
○ Spontaneous
○ Extremely spontaneous

7. Politeness

How polite/impolite is the facial expression?

Figure 2: Screenshot of the crowd-sourcing webpage.

Results

Data analysis was performed using standard statistical functions in MATLAB (R2014a, The MathWorks, Natick, USA).

First, we checked the correlations of evaluative dimensions within each setting. For this, ratings were averaged across actors, expressions, and participants and then correlated across dimensions. For the lab-based results (see Figure 3, left), we found that valence highly correlated with politeness, arousal with dominance and dynamics, persuasiveness with empathy and sincerity, and familiarity correlated with predictability. Naturalness correlated negatively with most other dimensions, since the question awarded posed expressions with high values. For the crowd-sourced results (see Figure 3, right), we similarly found that valence highly correlated with politeness, and arousal with dominance and dynamics. The similar patterns of inter-correlations already point towards consistent rating performance between the two settings.

To assess within-setting reliability we performed split-half correlations, correlating the responses across participants, but separately for each rating dimension. The split was repeated 1000 times, and we averaged the results. For the lab-based setting, reliability was very high overall with a median split-half correlation of r=.771 (see Figure 4, left). We found very high values for valence (r=.973), dynamics (r=.913), arousal (r=.900), politeness (r=.899), and audience (r=.855). Predictability (r=.387) and masculinity (r=.544) in contrast showed relatively low reliability. In the crowd-sourced setting, average correlation was also high (r=.727, Figure 4, right). Similarly, results for valence (r=.966), arousal (r=.859), politeness (r=.849), and dynamics (r=.797) showed these as highly reliable dimensions. Again, reliability was comparatively low for ratings of masculinity (r=.303) and predictability (r=.451).

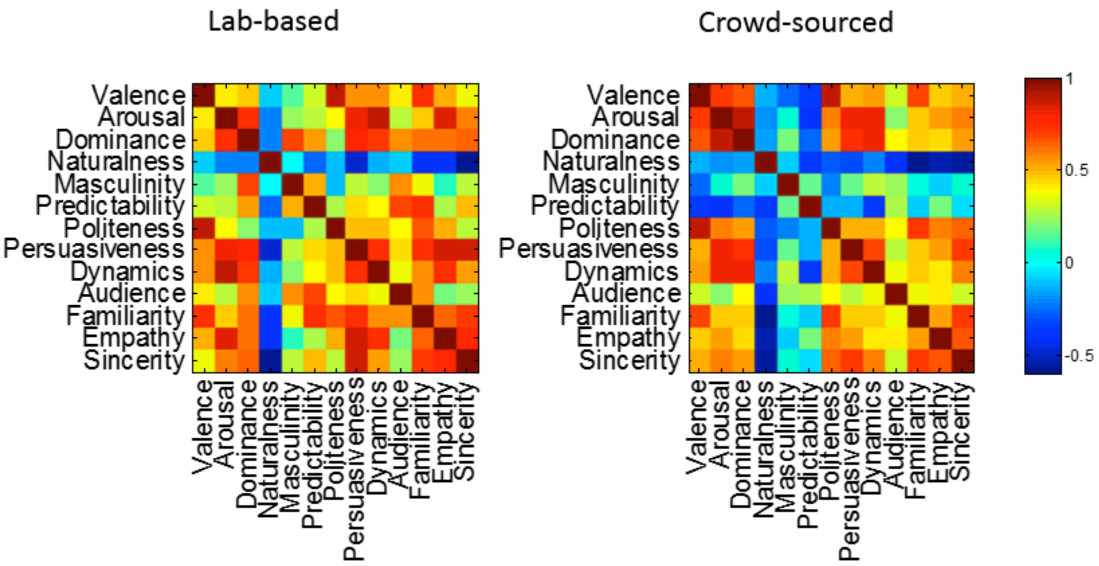


Figure 3: Correlations of evaluative dimensions within each setting (left: lab-based, right: crowd-sourced). Red indicates high and blue low correlation.

A Wilcoxon signed rank test for the two sets of correlation values found that lab-based correlations were significantly higher than those of the crowd-sourced setting (Z=2.201, p<.05). The overall difference in terms of effect size, however, was small (R=.169) indicating that although there was a significant difference (likely due to the large increase in between-participant noise because of the large number of participants), its impact was minor.

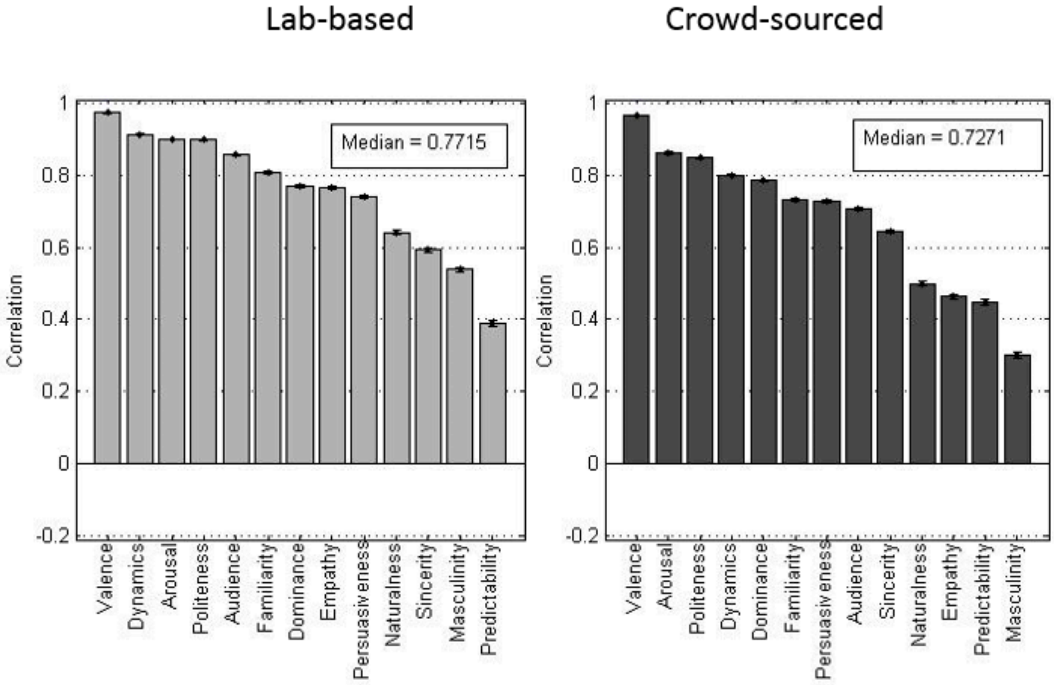


Figure 4: Reliability of ratings (as measured by split-half correlations) for each dimension. Left: lab-based setting, right: crowd-source setting.

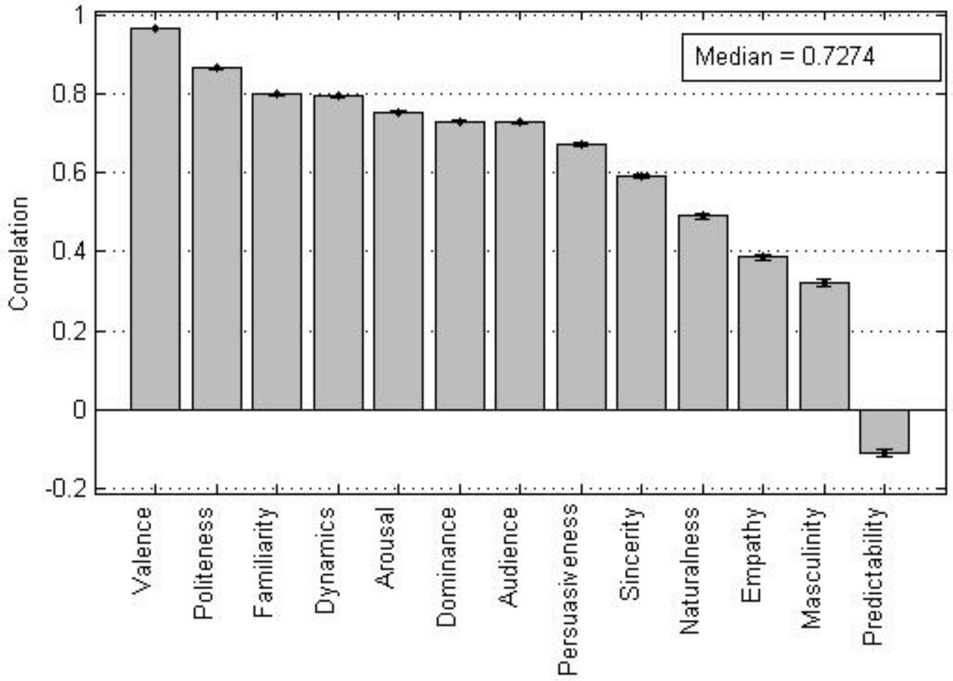


Figure 5: Across-setting correlations for each dimension.

Importantly, across-setting correlations for each dimension showed a highly similar rating pattern with a median correlation of r=.727 (see Figure 5). As may be expected, valence showed very high rating consistency (r=.966), but other dimensions such as politeness, familiarity, dynamics, and arousal also scored high across-setting correlations (r>.75). In accordance with their low reliability, predictability (r=-.106) and masculinity (r=.321) had low across-setting correlations. Importantly, for the aims of the present study, however, the overall correlations were high, hence confirming highly similar rating patterns in both the lab-based and the crowd-sourced setting.

Conclusion

Our results show that it is feasible to conduct large-scale rating experiments using crowd-sourcing stimuli. We found that valence and arousal as well as several other rating dimensions produced highly consistent ratings patterns in both lab-based and crowd-sourced settings. Masculinity and predictability in contrast were less reliable within each setting and accordingly also across settings.

Taken together, the ratings also paint a complex picture of how facial expressions are evaluated and which dimensions may be easily accessible for representing and processing the complex space of human communication. Future studies will use this data to conduct dimensionality analyses to investigate the space of human communicative expressions in more detail.

Importantly, for our goal, the present findings confirm previous studies about the potential and reliability of crowd-sourcing even with the complex video stimuli and ratings used here (e.g., Saunders et al. (2013)). In terms of resources used, the total time spent on running the lab-based experiment was 80 hours for 20 participants, whereas the crowd-sourcing experiment finished in 8 hours (using 423 participants), indicating a clear advantage for crowd-sourced solution.

It should be noted that although crowd-sourcing is effective and reliable when tested with general questions, differences between trained experts and crowd-sourced participants do exist: Nowak and Rüger (2010) compared image annotations of experts and crowd-sourced participants and showed that experts produced more consistent results, although the two groups did show good agreement. In our case, however, the rating task was supposed to tap into intuitive evaluative dimensions, hence needing no additional training or expertise.

Overall, we found that crowd-sourcing is a good alternative to lab-based experiments, enabling researchers to investigate cognitive dimensions of large numbers of visual stimuli easily and reliably.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2013R1A2A2A03068564) and the Brain Korea 21plus program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education.

References

Biele, C., & Grabowska, A. (2006). Sex differences in perception of emotion intensity in dynamic and static facial expressions. Experimental Brain Research, 171(1), 1-6.

Cameron, D. (2001). Working with spoken discourse. Sage.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. Computers in Human Behavior, 29(6), 2156-2160.

Castillo, S., Wallraven, C., & Cunningham, D. W. (2014). The semantic space for facial communication. Computer Animation and Virtual Worlds, 25(3-4), 223-231.

Coates, J. (2007). Talk in a play frame: More on laughter and intimacy. Journal of Pragmatics, 39(1), 29-49.

Cohen, I., Sebe, N., Garg, A., Chen, L. S., & Huang, T. S. (2003). Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and image understanding, 91(1), 160-187.

Cunningham, D. W. and C. Wallraven (2009). Dynamic information for the recognition of conversational expressions. Journal of Vision, 9(13), 7.

Drew P., & Heritage J. (1992). Analysing talk at work: An introduction. In P. Drew, & J. Heritage (Eds.), *Talk at work* (pp. 3-65). Cambridge: Cambridge University Press.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. Science, 164(3875): 86-88.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. Psychological science, 18(12), 1050-1057.

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's Mechanical Turk. Computers in Human Behavior, 29(4), 1749-1754.

Hsueh, P. Y., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing (pp. 27-35). Association for Computational Linguistics.

Kamachi, M., Bruce, V., Mukaida, S., Gyoba, J., Yoshikawa, S., & Akamatsu, S. (2001). Dynamic properties influence the perception of facial expressions. Perception, 30, 875-887.

Kaulard, K., Cunningham, D. W., Bülthoff, H. H., & Wallraven, C. (2012). The MPI Facial Expression Database—A validated database of emotional and conversational facial expressions. PloS one, 7(3), e32321.

Lee, H., Shin, A., Kim, B., & Wallraven, C. (2012). The KU facial expression database: a validated database of emotional and conversational expressions. In Proc. of Asian Pacific Conference on Vision, Incheon, Korea.

Mahmoud, M. M., Baltrusaitis, T., & Robinson, P. (2012). Crowdsourcing in emotion studies across time and culture. In Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia (pp. 15-16). ACM.

McDuff, D., Kaliouby, R. E., & Picard, R. W. (2012). Crowdsourcing facial responses to online videos. Affective Computing, IEEE Transactions on, 3(4), 456-468.

Morris, R. (2011). Crowdsourcing workshop: the emergence of affective crowdsourcing. In Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems. ACM.

Nowak, S., & Rüger, S. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In Proceedings of the international conference on Multimedia information retrieval (pp. 557-566). ACM.

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. Trends in cognitive sciences, 6(6), 261-266.

Sandbach, G., Zafeiriou, S., Pantic, M., & Yin, L. (2012). Static and dynamic 3D facial expression recognition: A comprehensive survey. Image and Vision Computing, 30(10), 683-697.

Saunders, D. R., Bex, P. J., & Woods, R. L. (2013). Crowdsourcing a normative natural language dataset: a comparison of Amazon Mechanical Turk and in-lab data collection. Journal of medical Internet research, 15(5).

Contact email: dilarad@korea.ac.kr