The Effect of Gender and Age Differences on the Recognition of Emotions from Facial Expressions

Daniela Schneevogt
University of Copenhagen
d.schneevogt@googlemail.com

Patrizia Paggio
University of Copenhagen
University of Malta
paggio@hum.ku.dk
patrizia.paggio@um.edu.mt

Abstract

Recent studies have demonstrated gender and cultural differences in the recognition of emotions in facial expressions. However, most studies were conducted on American subjects. In this paper, we explore the generalizability of several findings to a non-American culture in the form of Danish subjects. We conduct an emotion recognition task followed by two stereotype questionnaires with different genders and age groups. While recent findings (Krems et al., 2015) suggest that women are biased to see anger in neutral facial expressions posed by females, in our sample both genders assign higher ratings of anger to all emotions expressed by females. Furthermore, we demonstrate an effect of gender on the fear-surprise-confusion observed by Tomkins and McCarter (1964); females overpredict fear, while males overpredict surprise.

1 Introduction

Content in online social media is expressed not only in textual form, but also through pictures and videos. For example, YouTube has more than 1 billion users around the world, and it is estimated that 100 hours of video are uploaded every minute. Part of this content consists of video blogs where users express opinions about various topics. In order to mine the opinions that are expressed through images and videos, traditional text-based sentiment analysis must be complemented with similar techniques that are able to extract people's emotions and attitudes from the visual modality. Being able to extract emotions automatically, however, presupposes knowledge of how emotions are expressed and perceived. This paper focuses on two aspects of emotion understanding, i.e. whether gender and age play a role in the way people perceive emotions through facial expressions.

Several studies in the cognitive sciences have focused on studying people's perception and recognition of emotions in facial expression. The discussion about the relation between culture and emotions first started with Darwin in 1872 who argued that emotions and their expressions are universal (Darwin et al., 1998). Since then, an immense number of research studies on the universality of the basic emotions – anger, fear, disgust, happiness, sadness and surprise – has demonstrated that all healthy humans are able to recognize these emotions in images of human faces.

Recent studies also investigate possible differences in emotion recognition and find that subjects from different cultures indeed assign the same emotions, but show differences in perceived intensity and agreement level. Most recently, a study conducted with subjects from the American population shows an additional difference in emotion recognition based on gender (Krems et al., 2015). According to their research, women tend to see anger in other women's neutral facial expressions. Other studies indicate that these gender-specific differences decrease with advanced age (Calder et al., 2003; Mill et al., 2009).

Conducting a study only on an American sample raises the question of generalizability to other cultures. Therefore, our study was conducted to answer the following research questions: (1) For subjects with a different nationality than American, are there gender-specific differences in ratings of emotion expressions in human faces when offering a multiscalar rating scale? (2) Are these gender differences, if observed, comparable to the effects found by Krems et al. (2015)? (3) Do these differences change

Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 11–19, Osaka, Japan, December 12 2016.

with higher age? (4) Are the overall as well as the gender-specific ratings related to cultural and personal stereotypes? We chose to explore the first question by looking at Danish subjects, on the assumption that gender plays a different role in a Scandinavian culture.

We conduct an experiment including an emotion rating task with multiscalar rating scale and two questionnaires studying cultural stereotypes and personal beliefs. A selection of images from the NimStim Set of Facial Expressions (Tottenham et al., 2009) was used for the rating task. The questionnaires closely resemble the ones used by Plant et al. (2000) and are based on work by Fabes and Martin (1991). The experiment was conducted with 40 test subjects of two age groups.

2 Background

Over the years, a considerable number of studies have documented evidence for the existence of universal emotional expressions (e.g., Ekman (1994)). Various studies show that even across different cultures, humans are able to recognize the basic emotions of anger, fear, disgust, happiness, sadness and surprise in images of human faces. Recently, these studies expanded to investigate possible differences in emotion recognition across cultures and showed that subjects from different cultures indeed see the same emotions when presented with the same images, but show slight differences in the intensity ratings assigned to the respective emotion (Ekman et al., 1987; Matsumoto and Ekman, 1989; Biehl et al., 1997) and the precise level of agreement (Matsumoto et al., 2002; Russell, 1994; Biehl et al., 1997).

Biehl et al. (1997) and Martinez and Du (2012) also demonstrate the existence of different levels of agreement for emotion recognition, with happiness showing highest agreement, and disgust and fear the lowest. This confirms previous findings that showed fear to be the basic emotion most poorly recognized (Smith and Schyns, 2009; Biehl et al., 1997). A possible explanation is given by Tomkins and McCarter (1964) who suggest that emotions sharing similar expressive qualities – like fear and surprise – are most likely to be confused with each other. An alternative interpretation suggests that differences in the intensity assigned to an image could be based on low confidence (Biehl et al., 1997). Beaupré and Hess (2006) show that subjects are more confident when rating faces of in-group members and when rating expressions that are considered frequently expressed in their environment. Overall, subjects were found to be most confident in recognizing expressions of happiness (Beaupré and Hess, 2006), supporting the findings by Biehl et al. (1997).

Various research on cultural aspects of emotion recognition shows cultural differences in the level of agreement as well as the level of intensity assigned to the expressed emotions (Russell, 1994; Biehl et al., 1997; Elfenbein and Ambady, 2003). Further, Matsumoto (1991) found evidence for cultural display rules: While Japanese participants (collectivist culture) hide certain negative emotions when a person of higher status is present, American subjects (individualistic culture) openly show these emotions.

Several studies on gender differences in this field demonstrate a female advantage for decoding expressions of emotions. Hall and Matsumoto (2004) find that women are more accurate at identifying the correct pattern when rating emotions on a multiscalar rating scale, while this gender difference could not be observed for single choice tasks (Hall and Matsumoto, 2004). Moreover, women are believed to express sadness and fear more often than men, while men are believed to express anger more frequently (Fabes and Martin, 1991). Most recently, Krems et al. (2015) showed that women are biased to see anger in neutral female faces, whereas no such effect could be found for male faces or other emotions.

Plant et al. (2000) show that subjects behave in a stereotype-consistent manner when interpreting faces showing an ambiguous anger-sadness expression, rating women as more sad and less angry than men. Interestingly, even when shown unambiguous female anger expressions, participants rate these as a combination of anger and sadness, consistent with the subjects' stereotypes.

Studies on different age groups demonstrate that older subjects are less accurate at identifying emotions in facial expressions (Mill et al., 2009; Calder et al., 2003) and that this decline starts at 30 years of age for anger and sadness, and at 60 years of age for all other emotions (Mill et al., 2009). Calder et al. (2003) additionally show that performance on expressions of disgust improved for older participants.

3 Methodology

All data for our study was collected during May 2016. All experiments took place in a university laboratory and had the same protocol. Subjects were tested individually. First, subjects performed the emotion recognition task, and second, they filled in a Cultural Stereotype Questionnaire (CSQ) and a Personal Belief Questionnaire (PBQ). Afterwards, demographic data was collected. Two rounds of pilot experiments were run to eliminate technical and design-wise problems.

3.1 Stimuli selection

The data used in the emotion rating task was taken from the NimStim Set of Facial Expressions (Tottenham et al., 2009). The whole data set consists of 672 images of facial expressions posed by 43 actors. The emotions expressed are different versions of anger, disgust, fear, happiness, sadness and surprise, as well as neutral and calm expressions. From the total set of images, we first removed all images showing the calm facial expression as only the 6 basic emotions and neutral expressions were needed. Then, we selected a subset consisting of 112 images showing 7 distinct facial expressions by 16 actors – 8 males and 8 females. These actors were chosen such that for every actor all seven prototypical (Ekman and Friesen, 1978) expressions had an agreement score above 50 percent. In addition, 8 images were chosen for a practice phase in the beginning of each experiment session. The actors from the practice phase were not included in the final set of images.

3.2 Experimental setup and questionnaire design

The experiments were conducted using the open source software program OpenSesame (Mathôt et al., 2012) on an Apple MacBook Pro (13-inch) with OS X Yosemite (Version 10.10.3) installed. Each image (size: 506 x 650 pixels) appeared in the center of the screen for 5ms (the same duration as used in Beaupré and Hess (2005) and Beaupré and Hess (2006)). Each participant evaluated all 112 stimuli. The images were presented in a different random order for each of the 40 participants. After each image, a screen with a multiple rating scale for the 6 basic emotions (anger, disgust, fear, happiness, sadness, surprise) appeared. On screen, participants were instructed as follows: "Please indicate the extent to which you perceived each of the following emotions. 0=absent, 1=slight, 4=moderate, 8=strong". We based the design for the multiscalar rating in this task on the rating scale in Matsumoto (2005). Subjects then rated the intensity with which they perceived the 6 emotions in the previous image. Participants were allowed to either select one single emotion, several emotions, or to leave all emotions set to 0, indicating the absence of all 6 emotions and therefore a neutral expression.

After completing the rating task, participants were given two questionnaires studying subjects' stereotypes about the emotions of interest and additional questions on demographic details. Before starting the questionnaires, they were given oral instructions about the difference between them. Additional written instructions were given on screen for all parts of the experiment.

The two questionnaires we use in our study are short versions of stereotype endorsement questionnaires introduced by Fabes and Martin (1991) – a Cultural Stereotype Questionnaire (CSQ) and a Personal Beliefs Questionnaire (PBQ), and are based on the versions used by Plant et al. (2000). The CSQ studies cultural stereotypes about the frequency with which women and men experience and express emotions, while the PBQ investigates subjects' personal beliefs about this topic. Here, we study the same emotions examined in the emotion rating task – anger, disgust, fear, happiness, sadness and surprise.

For each emotion, subjects answered the following 4 questions on a scale from 1 (never) to 7 (very frequently) for the CSQ [1]:

(1) How often are men believed to experience ____?
(2) How often are men believed to express ____?
(3) How often are women believed to experience ____?
(4) How often are women believed to express ____?.

[1] We chose not to translate the questions into Danish since participants' proficiency in English was good enough to understand and answer the questions in English.

For each emotion in the PBQ, they were asked a slightly different version of each of the 4 questions, i.e. for the first question:

(1) How often do you believe men experience ____?

It was made clear both in the written and oral instructions that the first set of questions referred to general beliefs, while the second referred to the subject's own opinions. As the order in which subjects are presented the two questionnaires does not have an effect on the results (Plant et al., 2000), we present the questionnaires in the same order for all participants.

3.3 Participants

A total of 40 individuals participated in this study – 20 in the younger age group (10 male, 10 female) and 20 in the older age group (9 male, 11 female). All participants were Danish nationals born and raised in Denmark. In the younger age group, the mean age of the participants was 24.75 years (SD = 2.83) ranging from 18 to 31 years. This age group consisted of 16 university students from a variety of fields, as well as 2 high school students and 2 recent graduates. For the older age group, the mean age of the participants was 57.60 years (SD = 5.14), ranging from 50 to 69 years. Twelve of the older participants work at universities in a range of different positions (e.g. as professors, associate professors, senior researchers or research associates).

Out of all 40 participants, 33 reported having a university degree (Bachelor, Master or Doctorate degree) as the highest level of education. The remaining reported either a degree from primary school (3), highschool (2) or university college (2) as the highest. 19 participants reported a very high proficiency in English, 13 described their English as above average, 7 as average and only one person reported an English proficiency below average.

3.4 Analysis

To measure the accuracy with which subjects rated certain types of images (gender of expressor, emotion), we developed 2 different correctness measures. Both are developed on the assumption that there is always only one correct emotion, namely the emotion label given to each of the images in the data set (either neutral or one of the 6 emotions).

The first *correctness measure (C1)* looks at whether the highest rating given to an image corresponds to the correct emotion label. It was computed as follows for participant s rating an image i with a correct emotion label y:

$$C1(s,i,y) = \begin{cases} 100 & \text{if } y \neq neutral \wedge \arg\max_{e \in E} rating(s,i,e) = y \\ 100 & \text{if } y = neutral \wedge \sum_{e \in E} rating(s,i,e) = 0 \\ 0 & \text{otherwise} \end{cases}$$

Here, E is the set containing the 6 basic emotions. If the highest rating for a given image-participant pair is the rating for the correct emotion label, the value 100 is assigned, else the value 0. For neutral images, the value 100 is assigned if no emotions are given positive rating, else 0. Then, we calculate the average over all images per participant.

The second *correctness measure (C2)* considers the proportional rating of the correct emotion. It is based on the percentage of the rating given to the correct emotion label for an image over the sum of all ratings for that image. It is a real-valued variable (0%-100%) and was computed as follows: For every image-participant pair, we calculate the percentage of the total rating points given to the correct emotion label. Formally, $C2$ is calculated as follows:

$$C2(s,i,y) = \begin{cases} 100 \frac{rating(s,i,y)}{\sum_{e \in E} rating(s,i,e)} & \text{if } y \neq neutral \wedge \sum_{e \in E} rating(s,i,e) > 0 \\ 100 & \text{if } y = neutral \wedge \sum_{e \in E} rating(s,i,e) = 0 \\ 0 & \text{otherwise} \end{cases}$$

Afterwards, the average percentage for each participant over all stimuli is computed.

In addition to *C1* and *C2*, *emotion ratings* are used. Six real-valued variables (rating: 0-8) represent how highly rated an emotion was for a given image. The values were retrieved by performing the following calculations: For each emotion, the real value given to the respective emotion for each image-participant pair is collected. Then, the average rating over all images per participant is calculated for that emotion. Note that no value could be included for neutral, as neutral in our study was represented as the absence of ratings.

Finally, seven real-valued variables (0%-100%) represent – for each of the emotions – the percentages of the above described ratings over the sum of all ratings given for an image. These variables are the *emotion percentages*. To retrieve these values, we calculate the average percentage of the total given to the respective emotion for each participant over all stimuli. Values for neutral are calculated as in *C1* and *C2*.

For the questionnaires, six ordinal variables are used to refer to each of the six studied emotions – anger, disgust, fear, happiness, sadness and surprise.

Outlier detection was performed with SPSS' build-in function on both the emotion rating task and the questionnaire data. For the emotion rating task, two strong outliers were identified and excluded from further analysis. For the questionnaire data, no strong outliers were found. Therefore, all results we report in this paper are based on the analysis of data for 38 test subjects.

A Shapiro-Wilk test for normality showed that all dependent variables were normally distributed. Having established this, we ran the following ANOVAs, from which some results will be illustrated in Section 4:

Correctness (C1 and C2): 7×2×2×2 (emotion × poser gender × age group × subject gender) mixed ANOVA.
Emotion Ratings and Percentages: 7×2×2×2 (emotion × poser gender × age group × subject gender) mixed multivariate ANOVA.
Questionnaire Variables: 2×2×2×2×2 (subject gender × age group × target gender × questionnaire type × belief type) mixed multivariate ANOVA with 6 dependent variables, one per emotion.

All statistical analysis was carried out using SPSS Statistics, Version 22.0, and all post-hoc analysis was performed with a two-tailed *t*-test.

4 Results and Discussion

We begin by analyzing the subjects' performance on the emotion rating task. First, we show the mean values for both correctness measures for each emotion. The results can be seen in Table 1. To further illustrate the behavior of the subjects on the task, we plot in Figure 1 the average percentages of ratings assigned to each emotion.

As can be seen in Table 1, the subjects in our study showed a good overall performance on the emotion rating task when looking at the correctness results. An exception was participants' performance on neutral expressions, with correctness values of only 43.74% for both C1 and C2. This finding could be explained by task design: While all other emotional expressions could be given an intensity and rated along with other emotions, for neutrality this option was not given. As soon as a subject rated any other emotion with any intensity, the given image could not be rated as neutral anymore.

Interestingly, the results for correctness measure *C1* suggest a certain hierarchy: With a mean of 97.00%, subjects performed best at recognizing happiness compared to any other emotion, a result that is consistent with Beaupré and Hess (2006). Disgust, sadness, and surprise follow thereafter in the hierarchy, with respective means of 81.45%, 82.79%, and 86.17%. They are followed by anger at 71.67%, which is in turn followed by fear at 58.33%. At the bottom of the hierarchy are the neutral expressions. We see similar results for correctness measure *C2*, except for sadness and surprise for which the order was inverted. This hierarchical organization of the agreements on emotion ratings is consistent with the findings by Biehl et al. (1997).

Emotion	mean (C1)	mean (C2)
Anger	71.67%	70.26%
Disgust	81.45%	77.18%
Fear	**58.33%**	**60.65%**
Happiness	**97.00%**	**94.91%**
Sadness	82.79%	79.94%
Surprise	86.17%	77.43%
Neutral	43.74%	43.74%

Table 1: Mean values for expressed emotion on correctness measure C1 (Correct emotion ranked highest) and correctness measure C2 (Proportional rating of correct emotion).

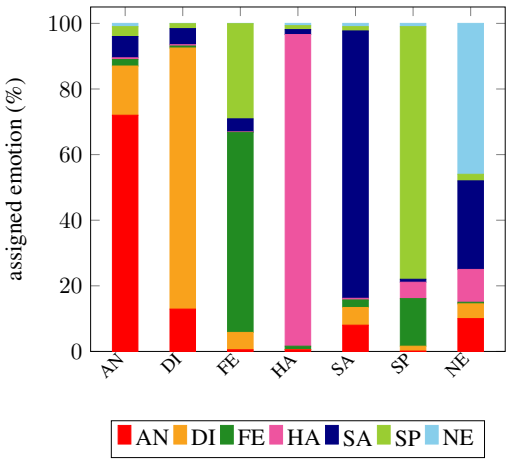

Figure 1: Average percentage of rating assigned to each emotion per expressor emotion.

Moreover, our results show that participants are outstandingly good at detecting happiness, with 97.00% and 94.91% for *C1* and *C2*, respectively. These numbers could be explained by the fact that happiness is the only strictly positive emotion in the set – the other emotions are all negative or ambiguous – and since distinguishing between positive and negative emotions is easier than discriminating between several options, happiness recognition is a relatively easy task.

Furthermore, as can be seen from Table 4, our subjects performed rather poorly at detecting emotional expressions of fear in human faces. In 41.67% of the cases where the expressed emotion was fear, another emotion was rated highest by the participants. Also, 39.35% of the rating points that were given to faces expressing fear were not given to fear but to one or more of the other emotions.

The emotion rating data shows that older subjects assign higher ratings of disgust to all emotions expressed. An interaction of emotion $\times$ age group with $p = .001$ ($F(6, 204) = 3.915$) for disgust was found. Post-hoc analysis shows that, on average, the older age group assigned a higher rating of disgust (6.21) to expressions of disgust than the younger age group (5.26). This effect was significant at $p = .008$. This outcome is consistent with the findings by Calder et al. (2003). No further age differences were found in our study.

To illustrate the effect of gender upon the rating task, we plot in Figures 2 and 3 the average percentages of ratings assigned to each emotion for each combination of subject gender and expressor gender.

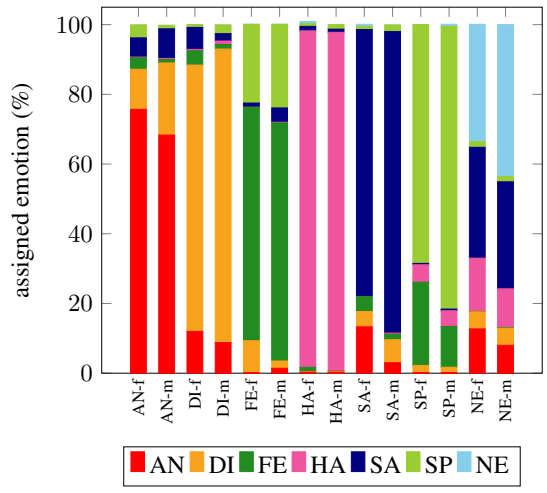

Figure 2: Average percentage of rating assigned to each emotion per expressor emotion and expressor gender for female subjects.

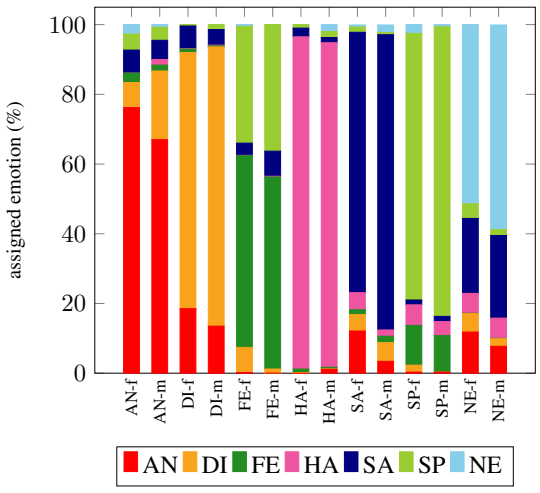

Figure 3: Average percentage of rating assigned to each emotion per expressor emotion and expressor gender for male subjects.

Contrary to what is found in Krems et al. (2015), in our study *both* genders assigned higher ratings of anger to *all* emotions expressed by female actors – on average 1.10 compared to 0.93 for male expressors ($p = .000$; $F(1, 34) = 15.537$). This can also be seen in Figures 2 and 3 – in both figures the value of anger is higher in the female bars than in the male bars (except for male subjects rating happiness, where a slight trend in the opposite direction is observed).

Furthermore, our analysis of fear and surprise expressions gives the following insight on the recognition of these two expressions: As in Tomkins and McCarter (1964), subjects confuse expressions of fear and surprise. However, they do this in gender-specific ways: Men predict more surprise, while women predict more fear. For the emotion ratings, an interaction was found for expressor gender $\times$ subject gender $\times$ emotion with $p = .004$ and $F(6, 204) = 3.300$. The post-hoc test shows that the mean rating given to fear by female subjects was 2.07, while the mean rating by male raters was 1.06. This effect is significant with $p = 0.012$. Females expressing surprise are rated as significantly more afraid if the rater is female.

For the emotion percentages, an interaction was found for emotion $\times$ subject gender. The significance values were $p = .013$ ($F(6, 204) = 2.760$) for fear and $p = .006$ ($F(6, 204) = 3.086$) for surprise. A post-hoc analysis shows that, for images of surprise, women assign higher percentages of fear than men do. For the female group, the mean percentage of fear assigned to surprise was 17.76%, men assigned 10.99% on average ($p = .044$). For images of fear, women assigned significantly less surprise than men did. Women on average assigned 23.07% of surprise to images of fear, whereas men on average assigned 35.03% surprise ($p = .041$).

Another interaction was detected for expressor gender $\times$ subject gender. For this interaction there is a significant effect at $p = .042$ ($F(1, 34) = 4.457$) for the percentage of fear rated. The post-hoc test reveals that on average images of women were rated as 14.19% fearful by female subjects, whereas male subjects on average assigned 10.37% fear to these images. This effect was significant with $p = .008$.

Moreover, an effect was found for the interaction of emotion $\times$ expressor gender $\times$ subject gender with $p = .005$ and $F(6, 204) = 3.163$. The percentage of ratings of fear given to female expressors posing surprise was significantly higher ($p = .012$) for female subjects (23.64%) than for male subjects (11.50%). Further, female subjects shown expressions of surprise assigned a significantly higher percentage of the ratings to fear if the expressor is female on average 23.64% to female expressors and 11.87% to male expressors ($p = .001$). It seems therefore that male subjects overcompensated by rating high values of surprise on both expressions of surprise and fear, while female subjects compensated by rating fear highly on expression of surprise and fear. This was especially true for female subjects rating female expressors.

The analysis of our questionnaire data also showed significant effects for fear, suggesting that subjects believe women are more prone to fear. For all emotions except anger, women were believed to experience and express these emotions more frequently than male targets do. The difference was especially large for fear – the average rating for male targets was 2.96, while it reached 4.95 for female targets. This was significant with $p = .000$ and $F(1, 34) = 46.289$. Plant et al. (2000) demonstrated that subjects prefer making predictions which are in accord with their own stereotypes, which could therefore serve as a possible explanation for our results.

Emotion	mean (express)	mean (experience)	p-value	$F(1, 34)$
Anger	3.88	4.46	.000	24.590
Disgust	3.48	3.92	.000	22.238
Fear	3.29	4.26	.000	42.219
Happiness	4.86	5.03	.169	1.975
Sadness	3.50	4.47	.000	32.870
Surprise	3.79	4.06	.054	3.991

Table 2: Means and significance values of the average rating by belief type and emotion.

Additionally, as can be seen in Table 2, our questionnaire data shows that all emotions are believed to be expressed less than they are experienced, suggesting that everyone is hiding emotions. A main effect of belief type was found for the 4 negative emotions anger, disgust, fear and sadness, all significant with $p = .000$. For all emotions, on average beliefs about experienced emotion are rated higher than beliefs about expressed emotions. The difference is especially high for anger, fear and sadness – three negative emotions – which indicates that people hide negative emotions more than positive emotions.

5 Conclusion

The goal of this study was to investigate gender and age differences in the recognition of emotions in facial expressions in a new cultural context. An emotion recognition study with subjects from the Danish population and two different age groups – followed by 2 questionnaires studying personal beliefs and cultural stereotypes – was conducted.

Consistent with relevant literature, the following findings were identified. First, emotions can be ordered by agreement hierarchy. Second, subjects show a good overall performance on the emotion rating task, especially for the emotion of happiness. Third, subjects perform the worst at detecting fear. And fourth, older subjects assign higher ratings of disgust.

In addition, we find that in our study *both* genders assign higher ratings of anger to *all* emotions expressed by female actors. This is in opposition to Krems et al. (2015), who found that women rated other womens' neutral faces as anger expressions. Furthermore, our subjects confuse expressions of fear and surprise, but in different, gender-specific manners: While men predict more surprise, women predict more fear. Our results show that, overall, gender plays an important role for the perception and recognition of emotions in facial expressions for Danish subjects, but in a different way than was found in Krems et al. (2015) among American subjects. Our results indicate that claims made about gender-specific differences in emotion recognition must take cultural factors into account.

This aspect could be studied in a more specific way in the future by investigating a possible correlation between measures of gender equality, perceived stereotypes and gender variation of emotion perception. Moreover, a separate experiment could be conducted to further examine the effects we have found concerning fear and surprise.

Another line of further investigation could deal with the difference between individual and collectivist cultures. Danish and American are in fact both considered individualist cultures. Interesting differences may arise when looking at data from a collectivist one, for example Japanese.

Finally, our study focuses on the recognition of emotions in static pictures. A much more complex, but certainly necessary domain in which emotions should be studied experimentally is that of videos, in which the understanding of emotions happens through the perception of multimodal expressions – facial expressions and speech.

Acknowledgements

We would like to thank the anonymous reviewers for their useful comments. Furthermore, we would like to thank all the participants who agreed to volunteer for our study.

References

Martin G. Beaupré and Ursula Hess. 2005. Cross-cultural emotion recognition among canadian ethnic groups. *Journal of Cross-Cultural Psychology*, 36(3):355–370.

Martin G. Beaupré and Ursula Hess. 2006. An ingroup advantage for confidence in emotion recognition judgments: The moderating effect of familiarity with the expressions of outgroup members. *Personality and Social Psychology Bulletin*, 32(1):16–26.

Michael Biehl, David Matsumoto, Paul Ekman, Valerie Hearn, Karl Heider, Tsutomu Kudoh, and Veronica Ton. 1997. Matsumoto and ekman's japanese and caucasian facial expressions of emotion (jacfee): Reliability data and cross-national differences. *Journal of Nonverbal Behavior*, 21(1):3–21.

Andrew J. Calder, Jill Keane, Tom Manly, Reiner Sprengelmeyer, Sophie Scott, Ian Nimmo-Smith, and Andrew W. Young. 2003. Facial expression recognition across the adult life span. *Neuropsychologia*, 41(2):195–202.

Charles Darwin, Paul Ekman, and Phillip Prodger. 1998. *The expression of the emotions in man and animals*. Oxford University Press, USA.

Paul Ekman and Wallace V. Friesen. 1978. *Manual for the facial action coding system*. Consulting Psychologists Press.

Paul Ekman, Wallace V. Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E. Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.

Paul Ekman. 1994. Strong evidence for universals in facial expressions: a reply to russell's mistaken critique.

Hillary Anger Elfenbein and Nalini Ambady. 2003. Universals and cultural differences in recognizing emotions. *Current Directions in Psychological Science*, 12(5):159–164.

Richard A Fabes and Carol Lynn Martin. 1991. Gender and age stereotypes of emotionality. *Personality and social psychology bulletin*, 17(5):532–540.

Judith A. Hall and David Matsumoto. 2004. Gender differences in judgments of multiple emotions from facial expressions. *Emotion*, 4(2):201.

Jaimie Arona Krems, Steven L. Neuberg, Gabrielle Filip-Crawford, and Douglas T. Kenrick. 2015. Is she angry?(sexually desirable) women see anger on female faces. *Psychological science*, 26(11):1655–1663.

Aleix Martinez and Shichuan Du. 2012. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *The Journal of Machine Learning Research*, 13(1):1589–1608.

Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes. 2012. Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods*, 44(2):314–324.

David Matsumoto and Paul Ekman. 1989. American-japanese cultural differences in intensity ratings of facial expressions of emotion. *Motivation and Emotion*, 13(2):143–157.

David Matsumoto, Theodora Consolacion, Hiroshi Yamada, Ryuta Suzuki, Brenda Franklin, Sunita Paul, Rebecca Ray, and Hideko Uchida. 2002. American-japanese cultural differences in judgements of emotional expressions of different intensities. *Cognition & Emotion*, 16(6):721–747.

David Matsumoto. 1991. Cultural influences on facial expressions of emotion. *Southern Journal of Communication*, 56(2):128–137.

David Matsumoto. 2005. Scalar ratings of contempt expressions. *Journal of Nonverbal Behavior*, 29(2):91–104.

Aire Mill, Jüri Allik, Anu Realo, and Raivo Valk. 2009. Age-related differences in emotion recognition ability: a cross-sectional study. *Emotion*, 9(5):619.

E. Ashby Plant, Janet Shibley Hyde, Dacher Keltner, and Patricia G. Devine. 2000. The gender stereotyping of emotions. *Psychology of Women Quarterly*, 24(1):81–92.

James A. Russell. 1994. Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies. *Psychological bulletin*, 115(1):102.

Fraser W Smith and Philippe G Schyns. 2009. Smile through your fear and sadness transmitting and identifying facial expression signals over a range of viewing distances. *Psychological Science*, 20(10):1202–1208.

Silvan S. Tomkins and Robert McCarter. 1964. What and where are the primary affects? some evidence for a theory. *Perceptual and motor skills*, 18(1):119–158.

Nim Tottenham, James W Tanaka, Andrew C. Leon, Thomas McCarry, Marcella Nurse, Todd A. Hare, David J. Marcus, Alissa Westerlund, BJ Casey, and Charles Nelson. 2009. The nimstim set of facial expressions: judgments from untrained research participants. *Psychiatry research*, 168(3):242–249.