Chapter 7
Across Cultures: A Cognitive and Computational Analysis of Emotional and Conversational Facial Expressions in Germany and Korea

Christian Wallraven, Dong-Cheol Hur, and Ahyoung Shin

Abstract Humans use a wide variety of communicative signals – among those, facial expressions play a key role in communicating not only emotional, but also more general, non-verbal signals. Here, we present results from a combined cognitive and computational analysis of emotional and conversational facial expressions in the context of cross-cultural research. Using two large databases of dynamic facial expressions, we show that both Western and Asian observers structure the interpretation space of a large range of facial expressions using the same two evaluative dimensions (valence and arousal). In addition, several computational experiments show the advantage of using graph-models for automatic recognition of facial expressions, since these models are able to capture the complex dynamics and inter-dependence of the movements of facial features in the face.

Keywords Facial expressions • Cross-cultural psychology • Emotions • Conversational expressions • Graph models

7.1 Introduction

Human communication can be divided into verbal and non-verbal signals. In the case of non-verbal signals, the human face plays a key role: the face itself conveys the person's identity and additional kinds of attributes such as attractiveness, intelligence, and trustworthiness, for example. Importantly, when the face starts to move, facial expressions are produced that convey information about one's feelings,

C. Wallraven (✉) • A. Shin
Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713,
Republic of Korea
e-mail: wallraven@korea.ac.kr

D.-C. Hur
Department of Computer Science and Engineering, Korea University, Seoul 136-713,
Republic of Korea

S.-W. Lee et al. (eds.), *Recent Progress in Brain and Cognitive Engineering*,
Trends in Augmentation of Human Performance, DOI 10.1007/978-94-017-7239-6_7

emotions, or intentions – the face starts to communicate. This “language of facial expressions” allows for rich and efficient interaction between people and forms one of the most important parts of non-verbal communication. It is therefore not surprising that there is a considerable amount of research about facial expressions in various fields, focusing on both cognitive aspects (that is, the investigation of how humans perceive, process, and use facial expressions) and computational aspects (that is, the investigation of how one may teach computers to understand and react to human facial expressions).

First, concerning cognitive aspects, an interesting debate in the field has been whether certain kinds of facial expressions may be “universal” signals of communication. For example, some cross-cultural studies have found evidence for highly robust interpretation and recognition of a certain group of six emotional facial expressions (the so-called “universal”, or basic facial expressions: anger, disgust, fear, happiness, sadness, and surprise). However, other studies have shown important differences even for these six expressions across different cultural backgrounds. In addition to this debate, another important aspect that has so far been largely neglected is that human communication consists of a much broader variety of signals that do not only transport emotional, but also cognitive and other socially-regulated intentions. In order to better understand how humans perceive and process facial expressions, detailed investigations of the perceptual and cognitive aspects of facial expression processing, taking into account the broad repertoire of expression signals as well as cross-cultural contexts, are necessary.

Secondly, concerning computational aspects of facial expression research, many frameworks for automatic and efficient human-computer-interaction have been proposed by computer vision researchers. So far, however, the available frameworks are not capable of recognizing more than the basic (six) emotional facial expressions. If we want to interpret the much larger range of general conversational expressions, we need to address some crucial obstacles: first, there is a large variability across individuals in expressing certain intentions, which represents a challenge for efficient modeling of expression categories. Second, conversational signals are often conveyed using highly subtle facial movements, which is another challenge for automatic facial feature tracking algorithms. Therefore, we need to develop novel computational frameworks that are capable of dealing with these issues such that we may interpret and process the full range of human communication.

7.2 Context in Brain and Cognitive Engineering

One of the core research foci of Brain and Cognitive Engineering is to use results from cognitive neuroscience to improve human-computer-interfaces and computer algorithms in general – depending on the domain used, one may call this biologically-, perceptually-, or cognitively-motivated computer science. Conversely, novel developments in machine learning and computer science can be used to increase our understanding of fundamental perceptual and cognitive processes in the brain. The present chapter offers a perspective on this research focus: we are first going to investigate the perceptual and cognitive aspects of facial expression processing in a cross-cultural context using state-of-the-art analysis methods. In a second step, we are then trying to design a computer vision system that takes into account some core aspects of the previously analyzed perceptual and cognitive data.

7.3 Previous Work

A series of studies conducted by Paul Ekman and colleagues in the 1960s and onwards have found evidence for the claim that certain facial expressions are "universal" across different cultures, that is, that recognizability and interpretability of these expressions is invariant across cultural contexts [1, 2]. However, recent research has cast doubts on the strong version of these statements and has shown that the concept of "universality" may be flawed, as results seem to be dependent on experimental and analysis methods [3–5]. Importantly, the aforementioned almost exclusively focused on emotional expressions only, although the full repertoire of facial expressions spans a much broader range of signals [6–9]. Indeed, relatively little is known about how we interpret and process conversational signals (such as a thoughtful or bored expression) or social signals (such as a wink, or a raised eyebrow). In addition to the cultural dependencies of emotional facial expressions, therefore the perceptual, cognitive, and cross-cultural aspects of more general facial expressions remain to be investigated [10–12].

The field of computational analysis of facial expressions has a rich history– albeit one focused again almost exclusively on the six basic facial expressions (for recent reviews, see [13, 14]). A recent recognition system achieving good recognition scores for the six emotional expressions by Kanaujia et al. [15] is based on an extended AAM (Active Appearance Model) that tracks the whole face. Many other frameworks are based on first detecting facial action units [16] (elementary muscle movements of the face) and then recognizing the six emotional expressions by detecting combinations of such facial action units [17]. These systems typically achieved the highest performance and are current state-of-the-art. Going beyond emotional expressions, Bousmalis et al. [18] tried to deal with two conversational expressions such as agreement and disagreement. In addition, McDuff et al. [19] developed an algorithm that is able to infer valence labels of continuous facial action sequences in unsegmented videos. However, most studies to date are based on rather constrained lab-settings and usually work only for a few kinds of facial expressions (mostly the six basic emotional expressions).

7.4 Database for Cross-Cultural Research

In order to conduct either cognitive neuroscience-related or computational research on conversational facial expressions, a suitable database is needed as a resource. Even when focusing on the six universal expressions, a large percentage of existing databases consists mainly of peak frames of expressions (i.e., static images) that do not contain dynamic movements. Indeed, several studies have shown that dynamic processing of facial expressions is different both in terms of behavioral (i.e., recognition accuracy [20]) and neuroimaging components [21]. Hence, the database needs to support dynamic stimuli in order to provide ecologically valid data.

Furthermore, when thinking about the broad range of human communication, there is a lack of databases containing conversational expressions. For this reasons, a few years ago we recorded the MPI facial expression database [22] that contains video sequences of both emotional and conversational facial expressions. Recently, we complemented this resource with a Korean equivalent: the KU facial expression database [23] that was developed with the exact same protocols than the German version. The MPI facial expression database has a total of 55 different facial expressions performed by 20 native Germans, whereas the KU facial expression database contains 55 plus 7 additional (=62) facial expressions performed by 20 native Koreans. The actors were recorded with three high-resolution video cameras yielding different points of view.

In order to ensure a good compromise between fully scripted (but potentially posed and unnatural) and unscripted (natural, but non-controlled) expressions, we employed a method-acting protocol during the recordings. For this, the experimenter read a developed scenario containing a short description of an event to the actor and asked them to imagine themselves in the scenario and to react accordingly. This process was repeated three times to yield three repetitions of each expression. The scenarios were designed to accommodate a large range of different emotional and conversational contexts. Importantly, the scenarios were designed with a conceptual hierarchy in mind: for example, there are many types of smile (pure smile, sad smile, reluctant smile, flirtatious smile, . . .) or many types of agreement (pure agreement, considered agreement, reluctant agreement, . . .). Indeed, for many types of expressions we were able to find a hierarchical structure. The full list of expressions and scenarios can be found in [22].

The resulting databases comprise two large (>20,000 video sequences), fully compatible datasets recorded in different cultural contexts. Examples for three different expressions are shown in Fig. 7.1.

After developing the databases, there was a validation step with both databases using German and Korean participants: for this, video sequences from each database were given to each group of participants, and each participant was asked to name the expressions corresponding to each video sequence using less than 4 words. Three independent raters then rated the answers as valid or invalid given the scenario descriptions. Using the most conservative criterion that a sequence is only rated as valid if it is approved by all three raters, on average, the MPI database and the KU database yielded 60 % and 57 % valid sequences, respectively. Using a less strict criterion of 2 out of 3 raters, we found validity scores of 71.5 % and 66.1 %, respectively.

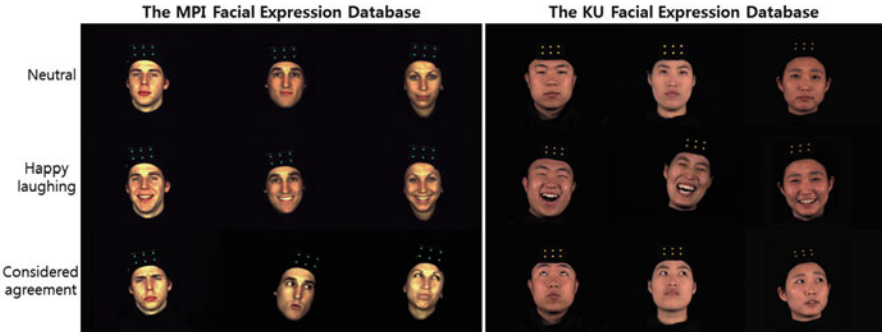

The MPI Facial Expression Database
The KU Facial Expression Database
Neutral
Happy laughing
Considered agreement

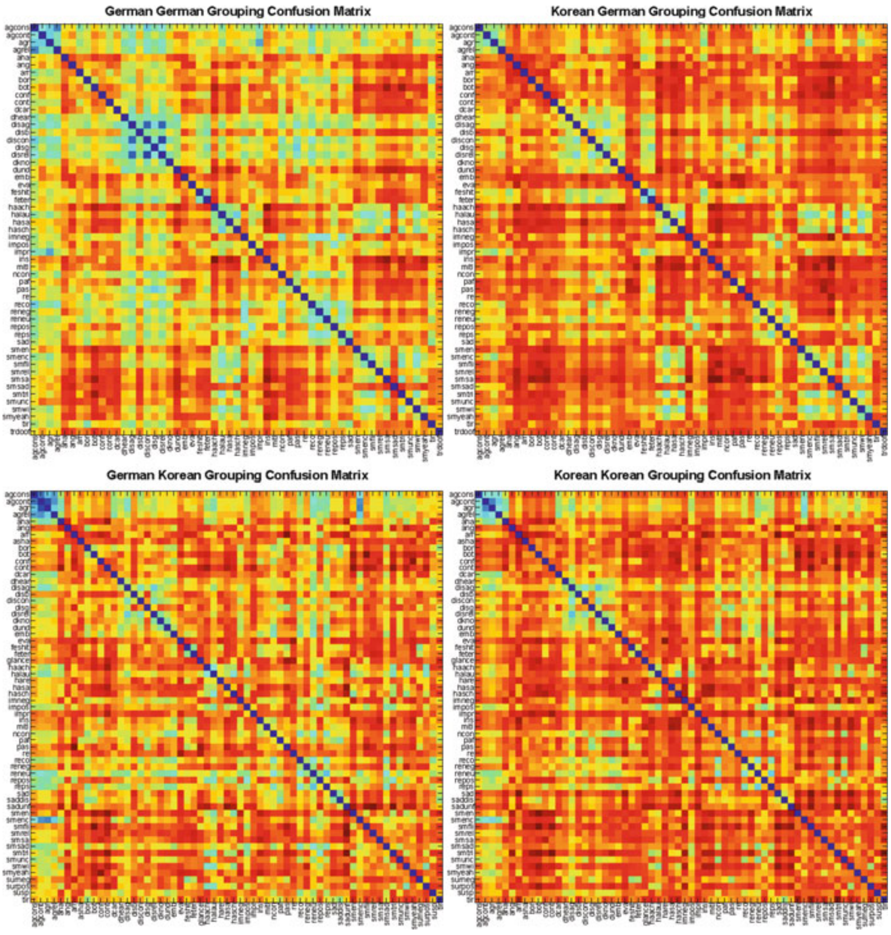
Fig. 7.1 Examples of the MPI (*left*) and KU (database) for three expressions. Note the considerable variation among individuals that is visible even in the static peak frames depicted here

7.4.1 Cognitive Study

We now turn to the first (cognitive) aspect of the present chapter in which we use the two databases to investigate the underlying dimensions of the complex space of emotional and conversational expressions in a cross-cultural context. For the experiments, 540 video sequences from the MPI facial expression database and 620 video sequences from the KU facial expression database were used as stimuli. Each group of stimuli contains expressions from 10 actors (MPI: 54 expressions of 10 actors, KU: 62 expressions of 10 actors). We conducted two fully crossed experiments across two countries, recruiting two participant groups in both Germany and Korea. For all experiments, we recruited only native German and Korean participants, where care was taken to control for exposure to the non-native cultural background (i.e., Asian/Korean for German participants, and Western/German for Korean participants). A total of 42 German participants and 44 Korean participants were recruited for this experiment.

The experiments consisted of a free-grouping task. Each group of participants received either 540 videos of the MPI facial expression database, or 620 videos of the KU facial expression database as video files in random order. Participants were then asked to group the expression sequences (i.e., to watch the video sequences and to move them into folders that they created one-by-one). There were no restrictions as to the number of clusters or the number of sequences in each cluster. In order to analyze the data, we generated confusion matrices for each of the four participant groups. Each confusion matrix tallies how often each expression was grouped with other expressions. With these matrices we then performed multidimensional scaling to identify the underlying topology and dimensionality of the resulting facial expression space.

The confusion matrices (see Fig. 7.2) showed similar structure for both databases as shown by the overall similar pattern: for example, the patterns for expressions belonging to the expression groups of "agreement" (expression labels starting with


German German Grouping Confusion Matrix

Korean German Grouping Confusion Matrix

German Korean Grouping Confusion Matrix

Korean Korean Grouping Confusion Matrix

Fig. 7.2 Confusion matrices for the four participant groups. *Blue* indicates similarly grouped expressions, whereas *red* indicates dissimilarly grouped expressions

"ag" in Fig. 7.2), "disagreement" (labels starting with "disag" in Fig. 7.2), and "thinking" (labels starting with "re" in Fig. 7.2) were seen as quite similar in all confusion matrices from German and Korean participants. In contrast, the data tended to yield more confusions for non-familiar cultural judgments: expressions belonging to the basic-level groups of "smiling" and "happiness" showed more confusion for cross-cultural (i.e., Korean-German Grouping or German-Korean Grouping in Fig. 7.2) than for within-cultural judgments.

Multidimensional scaling was then used to examine the first two dimensions of the low-dimensional embedding of the grouping data (see Fig. 7.3). Comparing the positions of the expressions located at the outsides of the space (e.g., for the KU facial expression database "eva" = evasive, "impr" = impressed,

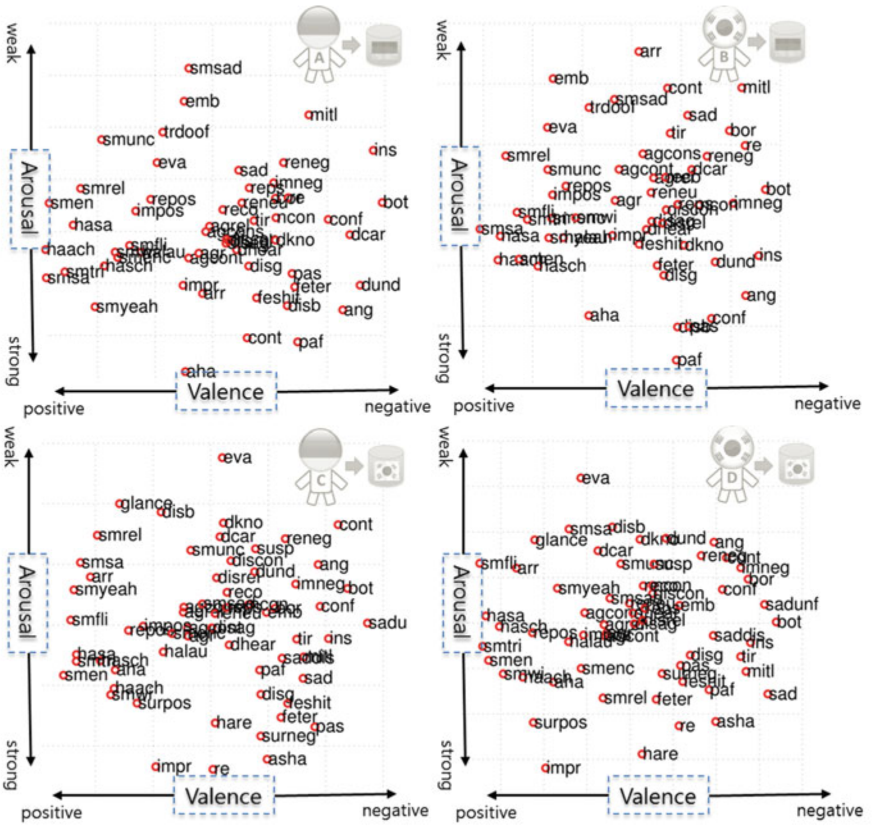


Fig. 7.3 Two-dimensional MDS solutions obtained from the four confusion matrices shown in Fig. 7.2

"smfli" = flirting smile, "bot" = bothered, and for the MPI facial expression database "emb" = embarrassed, "paf" = pain felt, "smsa" = sardonic smile, "bot" = bothered), we can see how similar the two reconstructed spaces are for each database. In addition, when comparing the KU and MPI databases, we can clearly see that expressions of the "smiling"-group (expressions starting with "sm") are located on the left side, whereas expressions such as anger ("ang") or bothered ("bot") are located on the right side of the plot. Hence this dimension recovered by the multidimensional scaling analysis corresponds to valence (positive-negative). A similar analysis reveals that the top-bottom dimension is that of arousal (weak-strong). Importantly, these dimensions are robustly recovered for both databases and both groups of participants. This shows that whereas there are differences between cultures (and to some degree, between databases), the overall structure of the space of facial expressions can be robustly explained by the two dimensions of valence and arousal.

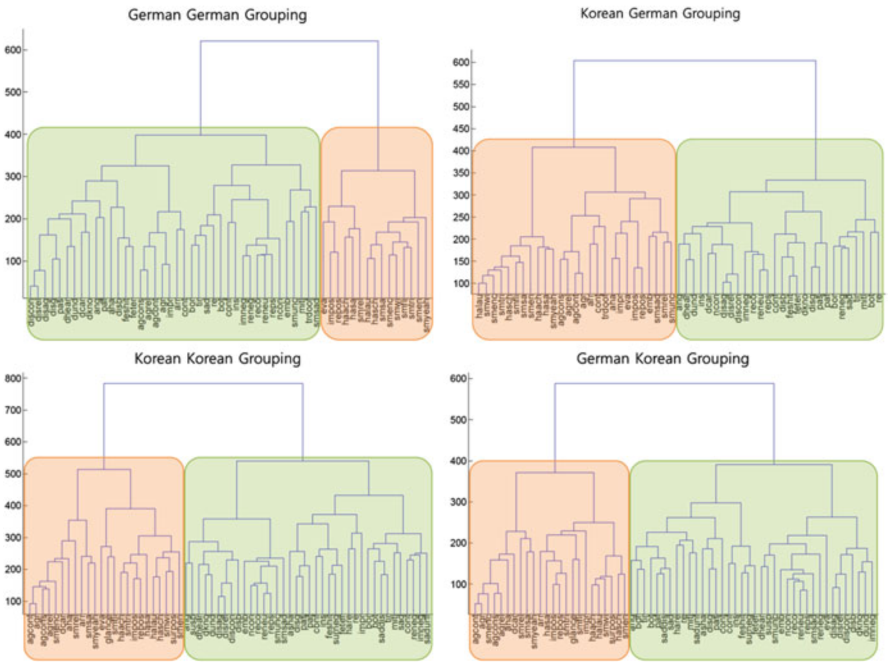


Fig. 7.4 Dendrograms from hierarchical clustering of the four participant groups. The expressions clustered in the two big groups mostly consist of positive expressions for the *red cluster* and negative expressions for the *green cluster*. The biggest difference appears for the German-German grouping, in which the “agree”-expressions (labels starting with “ag”) get grouped into the *green cluster*, which does not happen for the other grouping datasets

In addition, we employed bottom-up hierarchical clustering (using the Ward-criterion) to produce a clustering view of the data. Interestingly, when looking at the resulting four dendrograms, all expressions are divided into two big clusters at the first level for the four participant groups: these clusters include for the most part valence-positive (red) and valence-negative (green) expressions (see Fig. 7.4). Furthermore, the clusters on lower levels of the hierarchy re-produce our own conceptual hierarchy well: in most cases, agree, disagree, and thinking expressions receive their own cluster, for example. Hence, this clustering indirectly validates the hierarchical structure of conversational expressions that we used also during recording and design of the database.

7.4.2 Computational Study

Closer inspection of the two databases mentioned above shows that for many types of expressions there was considerable inter-person-variability – despite clear interpretability by human observers. Such variability will present a challenge for computational approaches. In addition, the number of categories (>50) is another issue that learning algorithms would need to deal with.

In the following, we present a computational recognition framework that tries to address these issues by using a powerful graphical sequence modeling approach: Latent Dynamic Conditional Random Fields (LDCRF). We train and test this modeling approach on the MPI database in the present chapter. For the computational experiments, we used expressions from 10 actors and took the first repetition as training data, and the second and third repetition of each expressions as testing data.

Importantly, we know that the structure of conversational facial expressions is hierarchical; for example, the expression of 'considered agreement' has two sub-expressions (considering and agreeing). In fact, these two sub-expressions can be shared across a wider range of expressions, since the "considering" part can also equally lead to a considered disagreement (another expression in the database) or simply stop without continuing (to yield thinking/considering, yet another expression in the database).

This observation also was the motivation for choosing LDCRFs for our task. Traditionally, Hidden Markov Models (HMMs) have often been used to model dynamic expressions. However, in order to predict the multiple categories of conversational expressions, Conditional Random Fields (CRFs) are more suitable because observations and latent states may follow conditional distributions. Furthermore, LDCRFs as an extension of CRFs are necessary since the hidden states (or latent factors) in LDCRFs are able to represent the required sub-expression dynamics discussed previously. Here, we compare the recognition performance of CRFs and LDCRFs in particular.

Both algorithms work on feature vectors extracted from the video sequences of the MPI facial expression database using the Computer Expression Recognition Toolbox [24]. The feature vectors include the intensity of 19 core facial action units as well as 3D-head rotations (yaw, pitch, roll). The result of tracking is used to automatically generate frame-wise labeling information. Since a manual annotation of each frame (as required by the CRF/LDCRF algorithms' training stage) would be a lot of work, we first set the intensity of the neutral expression (that by definition consisted of the first frame of each video sequence) based on the extracted feature vectors as the baseline. Each subsequent frame was then set as a non-neutral frame of the corresponding expression label, depending on a simple threshold difference to the neutral frame.

To compare the two algorithms (CRF and LDCRF), we chose to focus on six different expressions (considered agree, disagree, disgust, sad, I don't care, and happy laughing). Note that the sharing of sub-expressions here also would work in favor of the LDCRF, since sharing can of course also happen across the different actors. Accordingly, although the training time for LDCRF is longer the recognition rate is significantly higher than that of CRFs (88.6 % versus 77.1 %).

As an extension, we compared the previously mentioned human data with trained CRF models *on all expressions*. We compared confusion matrices that show how

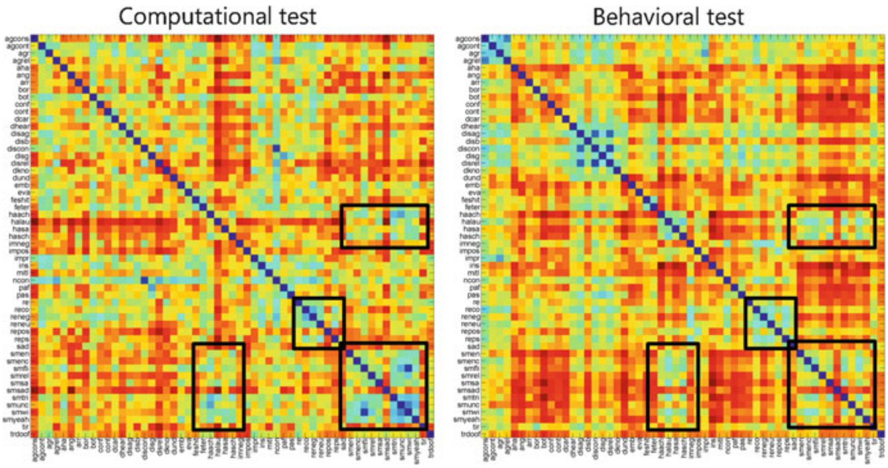


Fig. 7.5 Confusion matrices for computational (CRFs) and behavioral data (German-German grouping)

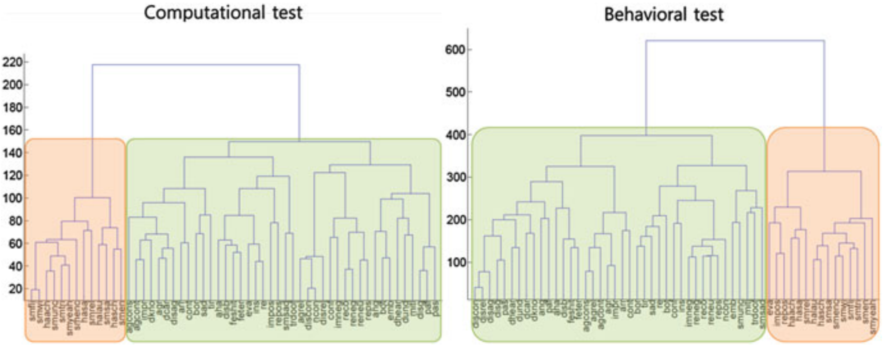


Fig. 7.6 Dendrograms for computational (CRFs) and behavioral test (German-German grouping). Note the similarity in the two larger (*red and green*) clusters between the two dendrograms – again, smiling expressions (expressions starting with "sm" or "ha") are split off first for both methods. Note, however, that clusters at lower levels of the hierarchy have differences. Cf. also Fig. 7.4 for human data

frequently expressions were confused with each other from both behavioral data and a simple computational test with the CRF models. The result is shown in Fig. 7.5. Although there are differences in the details, both matrices show a similar structure: for example, the clear block patterns in which "smiling" and "happy" categories are often confused is visible for both human and computational data (highlighted in Fig. 7.5).

This similarity is also visible in the hierarchical clustering, when comparing behavioral with computational data (see Fig. 7.6): again, the larger two clusters split off the smiling (valence-positive) expressions first. At the lower levels, however, differences start to appear: agree, and disagree expressions, for example, do not get grouped together in the computational clustering, whereas they clearly do in the behavioral data (see Fig. 7.4).

Overall, these results imply that relatively simple graphical models of computationally extracted features are able to replicate some broad-scale patterns of human performance.

7.5 Discussion

Using two large databases from two different cultural contexts, we first investigated cross-cultural perception of facial expressions – using expressions containing not only emotional, but also conversational aspects of facial expressions shown as dynamic data. We conclude that although expressions from a familiar background are more effectively grouped (i.e., less confused), the evaluative dimensions for both German and Korean cultural contexts are exactly the same, showing that cultural universals exist even in this complex space. The next step will consist of running rating experiments on a variety of conceptual scales to correlate the *implicit* dimensions obtained here (valence and arousal) to that of *explicit* judgments. Additional research will be conducted to extend this experiment to a larger participant base using crowd-sourcing to investigate different cultural backgrounds as well as dimensions of age.

For the computational study, we showed that since conversational expressions contain a hierarchical structure, modeling that takes into account this structure (LDCRF) shows a considerable advantage in recognition rates even on the smaller number of expressions tested here. In addition, we showed that the graphical models such as conditional random fields yield confusion patterns similar to those of human grouping on a broad scale. Future research will need to use more data from the full database (20 actors) to develop better models of facial expressions. Since such training is very costly at present with the extended CRF models, more efficient training algorithms will need to be developed as well to cope with the large amounts of data.

References

1. Ekman P (1994) Strong evidence for universals in facial expressions. Psychol Bull 115(2):268–287
2. Izard CE (1994) Innate and universal facial expressions: evidence from developmental and cross-cultural research. Psychol Bull 115(2):288–299
3. Russell JA (1994) Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychol Bull 115(1):102–141
4. Nelson NL, Russell JA (2013) Universality revisited. Emot Rev 5(1):8–15

5. Jack RE, Garrod OGB, Schyns PG (2013) Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. Curr Biol 24(2):187–192
6. Lee K-U, Khang HS, Kim K-T, Kim Y-J, Kweon Y-S, Shin Y-W, Liberzon I (2008) Distinct processing of facial emotion of own-race versus other-race. NeuroReport 19(10):1021–1025
7. Matsumoto D, Nakagawa S, Estrada A (2009) The role of dispositional traits in accounting for country and ethnic group differences on adjustment. J Pers 77(1):177–211
8. Jack RE, Blais C, Scheepers C, Schyns PG, Caldara R (2009) Cultural confusions show that facial expressions are not universal. Curr Biol 19(18):1543–1548
9. Jack RE, Garrod OGB, Yu H, Caldara R, Schyns PG (2012) Facial expressions of emotion are not culturally universal. Proc Natl Acad Sci 109(19):4–7
10. Schmidt KL, Cohn JF (2001) Human facial expressions as adaptations: evolutionary questions in facial expression research. Am J Phys Anthropol 33(S33):3–24
11. Elfenbein HA, Beaupré M, Lévesque M, Hess U (2007) Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. Emotion 7(1):131–146
12. McCarthy A, Lee K, Itakura S, Muir DW (2008) Gaze display when thinking depends on culture and context. J Cross Cult Psychol 39(6):716–729
13. Gatica-Perez D (2009) Automatic nonverbal analysis of social interaction in small groups: a review. Image Vis Comput 27(12):1775–1787
14. Vinciarelli A, Pantic M, Heylen D, Pelachaud C, Poggi I, D'Errico F, Schröder M (2012) Bridging the gap between social animal and unsocial machine: a survey of social signal processing. IEEE Trans Affect Comput 3(1):69–87
15. Kanaujia A, Metaxas D (2006) Recognizing facial expressions by tracking feature shapes. In: Proceedings – International conference on pattern recognition. Hongkong, vol 2, pp 33–38
16. Rivera J, Kreuz T (2009) Reading faces with conditional random fields, Technical report. Robotics Institute, Carnegie Mellon University, Pittsburgh
17. Chang KY, Liu TL, Lai SH (2009) Learning partially-observed hidden conditional random fields for facial expression recognition. In: 2009 IEEE computer society conference on computer vision and pattern recognition workshops. Miami, pp 533–540
18. Bousmalis K, Morency LP, Pantic M (2011) Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In: 2011 IEEE International conference on automatic face and gesture recognition and workshops. Santa Barbara, pp 746–752
19. McDuff D, El Kaliouby R, Kassam K, Picard R (2010) Affect valence inference from facial action unit spectrograms. In: 2010 IEEE computer society conference on computer vision and pattern recognition – workshops. San Francisco, pp 17–24
20. Cunningham DW, Wallraven C (2009) Dynamic information for the recognition of conversational expressions. J Vis 9(13):7.1–17
21. Bülthoff HH, Cunningham DW, Wallraven C (2011) Dynamic aspects of face processing in humans. In: Li SZ, Jain KA (eds) Handbook of face recognition. Springer, London, pp 571–576
22. Kaulard K, Cunningham DW, Bülthoff HH, Wallraven C (2012) The MPI facial expression database – a validated database of emotional and conversational facial expressions. PLoS One 7(3):e32321
23. Lee H, Shin A, Kim B, Wallraven C (2012) The KU facial expression database: a validated database of emotional and conversational expressions. In: Proceedings of Asian Pacific conference on vision. Incheon
24. Bartlett M, Littlewort G, Wu T, Movellan J (2008) Computer expression recognition toolbox (CERT). In: 2008 8th IEEE International conference on automatic face and gesture recognition. Amsterdam