3D Corpus of Spontaneous Complex Mental States

Marwa Mahmoud[1], Tadas Baltrušaitis[1], Peter Robinson[1], and Laurel D. Riek[2]

[1] Univeristy of Cambridge, United Kingdom
[2] University of Notre Dame, United States

Abstract. Hand-over-face gestures, a subset of emotional body language, are overlooked by automatic affect inference systems. We propose the use of hand-over-face gestures as a novel affect cue for automatic inference of cognitive mental states. Moreover, affect recognition systems rely on the existence of publicly available datasets, often the approach is only as good as the data. We present the collection and annotation methodology of a 3D multimodal corpus of 108 audio/video segments of natural complex mental states. The corpus includes spontaneous facial expressions and hand gestures labelled using crowd-sourcing and is publicly available.

1 Introduction

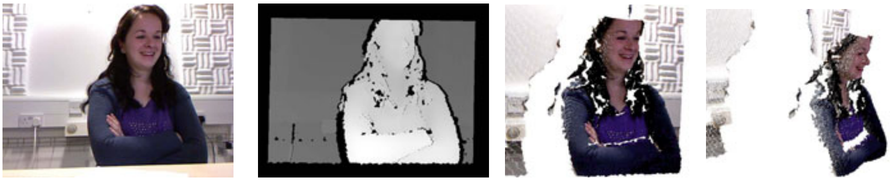
Human computer interaction could greatly benefit from automatic detection of affect from non-verbal cues such as facial expressions, non-verbal speech, head and hand gestures, and body posture. Unfortunately, the algorithms trained on currently available datasets might not generalise well to the real world situations in which such systems would be ultimately used. Our work is trying to fill this gap with a 3D multimodal corpus, which consists of elicited complex mental states. In addition, we are proposing hand-over-face gestures as a novel affect cue in affect recognition.

1.1 Motivation

There is now a move away from the automatic inference of the basic emotions proposed by Ekman [8] towards the inference of complex mental states such as attitudes, cognitive states, and intentions. The real world is dominated by neutral expressions [1] and complex mental states, with expressions of confusion, amusement, happiness, surprise, thinking, concentration, anger, worry, excitement, etc. being the most common ones [19]. This shift to incorporate complex mental states alongside basic emotions is necessary if one expects to build affect sensitive systems as part of a ubiquitous computing environment.

There is also a move towards analysing naturalistic rather than posed expressions, as there is evidence of differences between them [4]. In addition, even Action Unit [7] amplitude and timings differ in spontaneous and acted expressions [5]. These differences imply that recognition results reported on systems

S. D´Mello et al. (Eds.): ACII 2011, Part I, LNCS 6974, pp. 205–214, 2011.



(a) Colour image (b) Disparity map (c) Point cloud (d) Point cloud

Fig. 1. Point cloud visualisation from two angles (c)&(d) combining colour image (a) and disparity (inverse depth) map (b) of an image captured using Kinect

trained and tested on acted expressions might not generalise to spontaneous ones. Furthermore, this means that systems trained on current posed datasets would not be able to perform in tasks requiring recognition of spontaneous affect.

Hand-over-face gestures, a subset of emotional body language, are overlooked by automatic affect inferencing systems. Many facial analysis systems are based on geometric or appearance facial feature extraction or tracking. As the face becomes occluded, facial features are either lost, corrupted or erroneously detected, resulting in an incorrect analysis of the person's facial expression. Only a few systems recognise facial expressions in the presence of partial face occlusion, either by estimation of lost facial points [3,22] or by excluding the occluded face area from the classification process [6]. In all these systems, face occlusions are a nuisance and are treated as noise, even though they carry useful information.

Moreover, current availability of affordable depth sensors (such as the Microsoft Kinect) is giving easy access to 3D data, which can be used to improve the results of expression and gesture tracking and analysis. An example of such data captured using Kinect can be seen in Figure 1.

These developments are hindered by the lack of publicly available corpora, making it difficult to compare or reproduce results. Researchers cannot easily evaluate their approaches without an appropriate benchmark dataset.

1.2 Contributions

In order to address the issues outlined, we have collected and annotated a corpus of naturalistic complex mental states. Our dataset consists of 108 videos of 12 mental states and is being made freely available to the research community. The annotations are based on emotion groups from the Baron-Cohen taxonomy [2] and are based on crowd-sourced labels.

Moreover, we have analysed hand-over-face gestures and their possible meaning in spontaneous expressions. By studying the videos in our corpus, we argue that these gestures are not only prevalent, but can also serve as affective cues.

1.3 Related Work

In Table 1 we list several publicly available databases for easy comparison with our corpus Cam3D. This list is not exhaustive, for a more detailed one see Zeng

Table 1. Overview of similar databases

Properties	Cam3D	MMI[16]	CK+[14]	SAL[15]	BU-4DFE[21]	FABO[11]
3D	Y	N	N	N	Y	N
Modalities	S/F/U	F	F	S/F	F	F/U
Spontaneity	S	P/S	P/S	S	P	P
Number of videos	108	2894	700	10h	606	210
Number of subjects	7	79	210	24	100	24
Number of states	12	6	6	N/A	6	6
Emotional description	B/C	B	B	D/C	B	B,C

Modalities: S:speech, F:face, U:upper body, Spontaneity: S:spontaneous, P:posed, Emotional description: B:basic, C:complex, D:dimensional

et al. [25]. When compared in terms of modality and spontaneity all available datasets concentrate on some factors we are trying to address, while ignoring the others. MMI and CK+ corpora do not have upper body or hand gestures, SAL corpus consists of emotionally coloured interaction but lacks segments of specific mental states, while the FABO dataset contains only posed data.

Several 3D datasets of still images [24] and videos [21] of posed basic emotions already exist. The resolution of the data acquired by their 3D sensors is much higher than that available from Microsoft Kinect, but it is unlikely that such high quality imaging will be available for everyday applications soon.

2 Corpus

Care must be taken when collecting a video corpus of naturally-evoked mental states to ensure the validity and usability of the data. In the following sections, we will discuss our elicitation methodology, data segmentation and annotation.

2.1 Methodology

Elicitation of affective states. Data collection was divided into two sessions: interaction with a computer program and interaction with another person. Most available corpora are of emotions collected from single individuals or human-computer interaction tasks. However, dyadic interactions between people elicit a wide range of spontaneous emotions in social contexts under fairly controlled conditions [18]. Eliciting the same mental states during both sessions provides a comparison between the non-verbal expressions and gestures in both scenarios, especially if the same participant, stimuli, and experimental environment conditions are employed, and also enriches the data collected with different versions of expressions for the same affective state.

Our desire to collect multi-modal data presented a further challenge. We were interested in upper body posture and gesture as well as facial expressions to investigate the significance of hand and body gestures as important cues in non-verbal communication. Recent experiments have shown the importance of body language, especially in conditions where it conflicts with facial expressions [10]. Participants were not asked to use any computer peripherals during data collection so that their hands were always free to express body language.

Our elicitation methodology operated in four steps:

1. Choose an initial group of mental states.
2. Design an experimental task to induce them and conduct a pilot study.
3. Revise the list of the mental states induced according to the pilot results.
4. Validate the elicitation methodology after collecting and labelling the data.

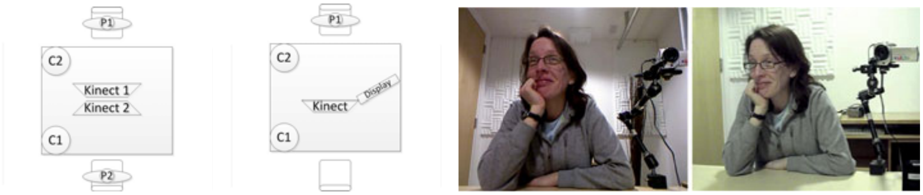
The first group of induced mental states were cognitive: *thinking, concentrating, unsure, confused* and *triumphant.* For elicitation, a set of riddles were displayed to participants on a computer screen. Participants answered the riddles verbally, with the option to give up if they did not know the answer. A second computer-based exercise was a voice-controlled computer maze. Participants were asked to traverse the maze via voice commands. In the dyadic interaction task, both participants were asked to listen to a set of riddles. They discussed their answers together and either responded or gave up and heard the answer from the speakers. In both tasks, the riddles' order was randomised to counter-balance any effect of the type of riddle on the results.

The second group of affective states were *frustrated* and *angry.* It was ethically difficult to elicit strong negative feelings in a dyadic interaction, so they were only elicited in the computer based session. During one attempt at the voice-controlled computer maze, the computer responded incorrectly to the participant's voice commands.

The third group included *bored* and *neutral.* It was also hard to elicit boredom intentionally in a dyadic interaction, so this was only attempted in the computer-based session by adding a 'voice calibration' task, where the participant was asked to repeat the words: left, right, up, down a large number of times according to instructions on the screen. Participants were also left alone for about three minutes after the whole computer task finished.

The last group included only *surprised.* In the computer-based session, the computer screen flickered suddenly in the middle of the 'voice calibration' task. In the dyadic interaction session, surprise was induced by flickering the lights of the room suddenly at the end of the session.

Experimental procedure. Data was collected in a standard experimental observation suite. A double mirror allowed experimenters to watch without disturbing the participants. In our experiment, a wizard-of-oz method was used for both the computer-based and dyadic interaction sessions. Participants knew at the beginning of the experiment that their video and audio were being recorded, but they did not know the actual purpose of the study. They were told that the experiment was for voice recognition and calibration. Not explaining the real objective of the experiment to participants in the first instance was essential for the data collection, to avoid having participants exaggerate or mask their expressions if they knew we were interested in their non-verbal behaviour.



(a) Dyadic task (b) Computer task (c) Examples of frames from Kinect and C2

Fig. 2. The layouts of the two parts of the data collection. P1 and P2 are the participants, C1 and C2 the HD cameras.

Participants. 16 participants (4 pilot and 12 non-pilot) were recruited through the university mailing lists and local message boards. The 12 non-pilot participants were 6 males and 6 females with age groups ranging from 24 to 50 years old (μ=27, σ=8). They were from diverse ethnic backgrounds including: Caucasian, Asian and Middle Eastern and with varied fields of work and study. All participants completed the two sessions: half the participants started with the computer-based task, while the other half started with the dyadic interaction task. Dyads were chosen randomly. Since cross-sex interactions elicit more non-verbal warmth and sexual interest than same-sex interactions [23], we chose same sex dyads to avoid this effect on the non-verbal expressions in the dyads. Half the participants (3 males, and 4 females) gave public consent for data distribution.

2.2 Data Acquisition

We used three different sensors for data collection: Microsoft Kinect sensors, HD cameras, and microphones in the HD cameras.

Figure 2a shows the layout for recording dyadic interactions. Two Kinect sensors and two HD cameras were used. The HD cameras each pointed at one participant and were used to record the voice of the other. In the case of the computer interaction task the camera layout is presented in Figure 2b. A Kinect sensor and an HD camera were facing the participant while one HD camera was positioned next to the participant and was facing away. The camera facing away was used to record the participant's voice.

Several computers were used to record the data from the different sensors, this required subsequent manual synchronisation.

The HD cameras provided 720 x 576 px resolution colour images at 25 frames per second. The recorded videos were later converted to 30 frames per second to simplify synchronisation with Kinect videos.

The Kinect sensor provides a colour image and a disparity map, which is the inverse of depth values, at 30 frames per second. The sensor uses structured infrared light and an infrared camera to calculate 640 x 480 px 11-bit disparity map. An additional camera provides a 640 x 480 colour image.

2.3 Segmentation and Annotation

After the initial data collection, the videos were segmented. Each segment showed a single event such as a change in facial expression, head and body posture movement or hand gesture. This increases the value of the annotation compared with cutting the whole video into equal length segments [1].

Video segments were chosen and annotated using ELAN [13]. From videos with public consent, a total of 451 segments were collected. The mean duration is 6 seconds (σ=1.2). For subsequent analysis, each video segment was annotated with the type of the task and interaction. In addition, we encoded hand-over-face gestures (if any) in terms of: hand shape, action, and facial region occluded. From the non-public videos recorded, only hand-over-face gestures (120 segments) were segmented and encoded to be included in subsequent hand gestures analysis.

Labelling was based on context-free observer judgment. Public segments were labelled by community crowd-sourcing, which is fast, cheap and can be as good as expert labelling [20]. The sound was low-pass filtered to remove the verbal content of speech. Video segments were displayed randomly through a web interface and participants were asked to give a 'word' describing the emotional state of the person in the video. Free-form input was used rather than menus in order not to influence the choice of label. In addition, an auto-complete list of mental states was displayed to avoid mis-spelling. The list was based on the Baron-Cohen taxonomy of emotions [2] as it is an exhaustive list of emotional states and synonyms (1150 words, divided into 24 emotion groups and 412 emotion concepts with 738 synonyms for the concepts). In total 2916 labels from 77 labellers were collected (μ=39). Non-public video segments were labelled by four experts in the field and segments with less than 75% agreement were excluded.

We decided to use categorical labels rather than continuous ones such as PAD (pleasure, arousal, dominance) scales because we used naive labellers. Dimensional representation is not intuitive and usually requires special training [25], which would be difficult when crowd-sourcing.

3 Data Analysis

3.1 Validation

Out of the 451 segmented videos we wanted to extract the ones that can reliably be described as belonging to one of the 24 emotion groups from the Baron-Cohen taxonomy. From the 2916 labels collected, 122 did not appear in the taxonomy so were not considered in the analysis. The remaining 2794 labels were grouped as belonging to one of the 24 groups plus *agreement*, *disagreement*, and *neutral*.

Because raters saw a random subset of videos not all of them received an equal number of labels. We did not consider the 16 segments that had fewer than 5 labels. To filter out non-emotional segments we chose only the videos that 60% or more of the raters agreed on. This resulted in 108 segments in total. As the average number of labels per video was 6 the chance of getting 60% agreement by chance is less than 0.1%. The most common label given to a video segment

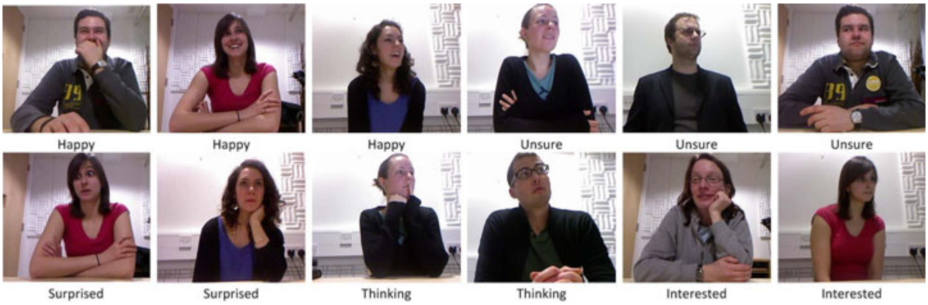


Fig. 3. Example of still images from the dataset

was considered as the ground truth. Examples of still images from the labelled videos can be seen in Figure 3.

We validated the labelling of the selected videos using Fleiss's Kappa (κ) [9] measure of inter-rater reliability. The resulting $\kappa = 0.45$ indicates moderate agreement. This allows us to dismiss agreement by chance, and be confident in annotating the 108 segments with the emotional group chosen by the annotators.

Alternatively, if we were to choose a higher cutoff rate of 70% (56 videos) or 80% (40 videos), instead of 60% we would get $\kappa = 0.59$ and $\kappa = .67$ respectively, reaching substantial agreement. Although this would lead to fewer videos in our corpus, those videos might be seen as better representations of the mental states. We reflect this in our corpus by reporting the level of agreement per segment. Probabilistic systems can benefit from knowing the uncertainty in the ground truth and exploit that in classification.

Furthermore, we wanted to estimate inter-rater agreement for specific mental states in the resulting 108 segments. Expressions of basic emotions of *happy* ($\kappa = 0.64$) and *surprised* ($\kappa = 0.7$), had higher levels of agreement than complex mental states of *interested* ($\kappa = 0.32$), *unsure* ($\kappa = 0.52$), and *thinking* ($\kappa = 0.48$). For this analysis we only consider the expressions with no fewer than 5 representative videos.

We also wanted to see how successful certain elicitation methods were at generating certain naturalistic expressions of affect. Most of the affective displays came from the riddles both in computer and dyadic tasks. They were successful at eliciting *thinking* (26) and *unsure* (22), with some *happy* (14), *surprised* (3), *agreeing* (5), and *interested* (2). The longer and more complicated maze was successful at eliciting *interest* (4) and *thinking* (1). The third maze managed to elicit a broader range of expressions, including *surprised* (1), *sure* (1), *unsure* (1) and *happy* (3). This was somewhat surprising as we did not expect to elicit happiness in this task. There is some evidence [12] of people smiling during frustration which might explain perception of happiness by labellers.

3.2 Analysis of Hand-Over-Face Gestures

In *The Definitive Book of Body Language*, Pease and Pease [17] attempt to identify the meaning conveyed by different hand-over-face gestures. Although

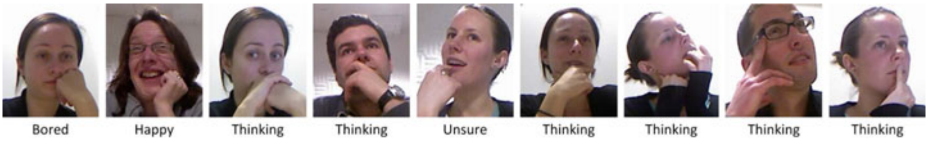


Fig. 4. Different hand shape, action and face region occluded are affective cues in interpreting different mental states

they suggest that different positions and actions of the hand occluding the face can imply different affective states, no quantitative analysis has been carried out. Using collected video segments, we have analysed hand-over-face gestures included in the videos in terms of the hand shape and its action relative to face regions. In the 451 initial public segments collected, hand-over-face gestures appeared in 20.8% of the segments (94 segments), with 16% in the computer-based session and 25% in the dyadic interaction session. Participants varied in how much they gestured, some had a lot of gestures while others only had a few.

Looking at the place of the hand on the face in this subset of the 94 hand-over-face segments, the hand covered upper face regions in 13% of the segments and lower face regions in 89% of them, with some videos having the hand overlapping both upper and lower face regions. This indicates that in naturalistic interactions hand-over-face gestures are very common and that hands usually cover lower face regions, especially chin, mouth and lower cheeks, more than upper face regions.

We analysed the annotated corpus of the 108 video segments in addition to the expert-labelled private segments. Total hand-over-face segments studied were 82. Figure 4 presents examples of labelled segments of hand-over-face gestures.

In the publicly labelled set, hand-over-face gestures appeared in 21% of the segments. Figure 5 shows the distribution of the mental states in each category of the encoded hand-over-face gestures. For example, index finger touching face appeared in 12 *thinking* segments and 2 *unsure* segments out of a total of 15 segments in this category. The mental states distribution indicates that passive hand-over-face gestures, like leaning on the closed or open hand, appear in different mental states, but not in cognitive mental states. On the other hand, actions like stroking, tapping and touching facial regions - especially with index finger - are all associated with cognitive mental states, namely *thinking* and *unsure*. Thus, hand shape and action on different face regions can be used as a novel cue in interpreting cognitive mental states.

4 Discussion

We have described the collection and annotation of a 3D multi-modal corpus of naturalistic complex mental states, consisting of 108 videos of 12 mental states. The annotations are based on crowd-sourced labels. Over six hours of data was collected, but only generated 108 segments of meaningful affective states, which highlights the challenge of collecting large naturalistic datasets. Analysing our

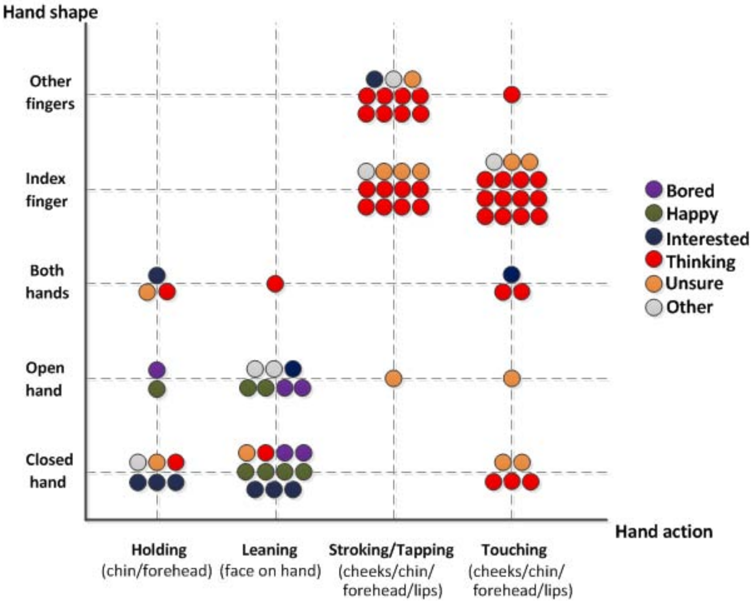


Fig. 5. Encoding of hand-over-face shape and action in different mental states. Note the significance of the index finger actions in cognitive mental states.

corpus, we noticed the potential of hand-over-face gestures as a novel modality in facial affect recognition. Our mental states elicitation methodology was successful; therefore, future work will include adding more data to our corpus. This will allow further exploration spontaneous gestures and hand-over-face cues. Furthermore, we are exploring the use of depth in automatic analysis of facial expressions, hand gestures and body postures.

Acknowledgment. We acknowledge funding support from Yousef Jameel Scholarships, EPSRC and Thales Research and Technology (UK). We would like to thank Lech Swirski and Andra Adams for their help in video recording and labelling.

References

1. Afzal, S., Robinson, P.: Natural affect data - collection & annotation in a learning context. In: ACII, pp. 1–7. IEEE, Los Alamitos (2009)
2. Baron-Cohen, S., Golan, O., Wheelwright, S., Hill, J.: Mind Reading: The Interactive Suide to Emotions (2004)
3. Bourel, F., Chibelushi, C., Low, A.: Robust facial expression recognition using a state-based model of spatially-localised facial dynamics. In: IEEE AFGR (2002)
4. Cowie, R.: Building the databases needed to understand rich, spontaneous human behaviour. In: AFGR, pp. 1–6. IEEE, Los Alamitos (2008)

5. Duchenne, G., Cuthbertson, R.: The mechanism of human facial expression. Cambridge Univ. Press, Cambridge (1990)
6. Ekenel, H., Stiefelhagen, R.: Block selection in the local appearance-based face recognition scheme. In: CVPRW, pp. 43–43. IEEE, Los Alamitos (2006)
7. Ekman, P., Friesen, W.: Manual for the Facial Action Coding System. Consulting Psychologists Press, Palo Alto (1977)
8. Ekman, P., Friesen, W.V., Ellsworth, P.: Emotion in the Human Face, 2nd edn. Cambridge University Press, Cambridge (1982)
9. Fleiss, J., Levin, B., Paik, M.: Statistical Methods for Rates and Proportions. Wiley, Chichester (2003)
10. de Gelder, B.: Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. Phil. Trans. of the Royal Society B 364(1535), 3475 (2009)
11. Gunes, H., Piccardi, M.: A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In: ICPR, vol. 1, pp. 1148–1153. IEEE, Los Alamitos (2006)
12. Hoque, M.E., Picard, R.W.: Acted vs. natural frustration and delight: Many people smile in natural frustration. In: IEEE AFGR (2011)
13. Lausberg, H., Sloetjes, H.: Coding gestural behavior with the NEUROGES-ELAN system. Behavior research methods (2009), http://www.lat-mpi.eu/tools/elan/
14. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: CVPRW, pp. 94–101. IEEE, Los Alamitos (2010)
15. McKeown, G., Valstar, M., Cowie, R., Pantic, M.: The SEMAINE corpus of emotionally coloured character interactions. In: ICME, pp. 1079–1084. IEEE, Los Alamitos (2010)
16. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: IEEE Conf. Multimedia and Expo, p. 5. IEEE, Los Alamitos (2005)
17. Pease, A., Pease, B.: The definitive book of body language, Bantam (2006)
18. Roberts, N., Tsai, J., Coan, J.: Emotion elicitation using dyadic interaction tasks. In: Handbook of Emotion Elicitation and Assessment, pp. 106–123 (2007)
19. Rozin, P., Cohen, A.B.: High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. Emotion 3(1), 68–(2003)
20. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.: Cheap and fast-but is it good?: evaluating non-expert annotations for natural language tasks. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing, pp. 254–263. Association for Computational Linguistics (2008)
21. Sun, Y., Yin, L.: Facial expression recognition based on 3D dynamic range model sequences. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 58–71. Springer, Heidelberg (2008)
22. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. IEEE PAMI, 1683–1699 (2007)
23. Weitz, S.: Sex differences in nonverbal communication. Sex Roles 2, 175–184 (1976)
24. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: AFGR, pp. 211–216 (2006)
25. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. TPAMI 31(1), 39–58 (2009)