Image and Vision Computing xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database ☆

Xing Zhang [a,*], Lijun Yin [a,*], Jeffrey F. Cohn [b], Shaun Canavan [a], Michael Reale [a], Andy Horowitz [a], Peng Liu [a], Jeffrey M. Girard [b]

[a] *State University of New York at Binghamton, United States*
[b] *University of Pittsburgh, United States*

ARTICLE INFO

Article history:
Received 13 June 2013
Received in revised form 17 May 2014
Accepted 26 June 2014
Available online xxxx

Keywords:
3D facial expression
FACS
Spontaneous expression
Dynamic facial expression database

ABSTRACT

Facial expression is central to human experience. Its efficiency and valid measurement are challenges that automated facial image analysis seeks to address. Most publically available databases are limited to 2D static images or video of posed facial behavior. Because posed and un-posed (aka "spontaneous") facial expressions differ along several dimensions including complexity and timing, well-annotated video of un-posed facial behavior is needed. Moreover, because the face is a three-dimensional deformable object, 2D video may be insufficient, and therefore 3D video archives are required. We present a newly developed 3D video database of spontaneous facial expressions in a diverse group of young adults. Well-validated emotion inductions were used to elicit expressions of emotion and paralinguistic communication. Frame-level ground-truth for facial actions was obtained using the Facial Action Coding System. Facial features were tracked in both 2D and 3D domains. To the best of our knowledge, this new database is the first of its kind for the public. The work promotes the exploration of 3D spatiotemporal features in subtle facial expression, better understanding of the relation between pose and motion dynamics in facial action units, and deeper understanding of naturally occurring facial action.

1. Introduction

Research on computer-based facial expression and affect analysis has intensified since the first FG conference in 1995. The resulting advances have made the emerging field of affective computing possible. The continued development of emotion-capable systems greatly depends on access to well-annotated, representative affective corpora [13]. A number of 2D facial expression databases have become available (e.g., [1,2,7,8,16]), as well as some with 3D imaging data (e.g., [9,14,15,24,25,45]). Although some systems have been successful, performance degrades when handling expressions with low intensity appearance, large head rotation, subtle skin movement, and/or lighting changes with varying postures.

Due to the limitations of describing facial surface deformation when 3D features are evaluated in 2D space, 2D images with a handful of features may not accurately reflect the authentic facial expressions (e.g., in-depth motion of 3D head pose, 3D wrinkles and skin extrusion in the areas of the cheek, forehead, glabella, nasolabial, and crow's feet).

Another problematic issue is that facial action units (AUs) can occur in more than 7000 different complex combinations [17], causing bulges and various in- and out-of-image-plane movements of permanent facial features, negative emotions from view of the left hemiface, and mouth extrusions that are difficult to detect in a 2D plane. Three-dimensional dynamic surface analysis and tracking will be important for those facial expressions for which 2D motion information is not sufficient.

Because the face is a 3D object and many communicative signals involve changes in depth and head rotation, inclusion of 3D information is an important addition. Another major limitation of existing databases is that most have only posed or acted facial behavior, and thus the data are not representative of spontaneous affective expression, which may differ in timing, complexity, and intensity from posed expression [22]. No currently available dataset contains *dense*, *dynamic*, *3D* facial representations of *spontaneous* facial expression with anatomically-based (FACS) *annotation* [36].

Currently, most approaches to automatic facial expression analysis attempt to recognize a set of prototypic emotional expressions (e.g., anger, disgust, fear, happiness, sadness, and surprise) [3,5,13]. Many studies about emotion used "acting" or "emotion portrayals" in a restricted sense by recording subjects who are instructed to express single-label emotions, sometimes using scripts [6]. However, the resulting posed and exaggerated facial actions may occur only rarely in daily life [4].

Because posed and un-posed (aka "spontaneous") facial expressions differ along several dimensions [32], including complexity (especially with respect to segmentation), well-annotated video of un-posed facial behavior is needed. Moreover, as noted above, because the face is a three-dimensional deformable object, a 3D video archive would be

☆ This paper has been recommended for acceptance by Qiang Ji.

* Corresponding authors.
E-mail addresses: xzhang7@cs.binghamton.edu (X. Zhang), lijun@cs.binghamton.edu (L. Yin).

http://dx.doi.org/10.1016/j.imavis.2014.06.002
0262-8856/© 2014 Elsevier B.V. All rights reserved.

Please cite this article as: X. Zhang, et al., BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database, Image Vis. Comput. (2014), http://dx.doi.org/10.1016/j.imavis.2014.06.002

especially important. Two-dimensional databases, such as RU-FACS [23] or Cohn–Kanade [2], are insufficient. The CMU Multi-PIE database [34], 3D dynamic AU database [35], Bosphorus database [9], KDEF [33], BU 3D facial expression databases [14,15], and ICT-3DRFE database [24] begin to address the need for 3D (or multi-view) data but are limited to posed facial behavior.

Recent efforts to collect, annotate, and analyze spontaneous facial expression for community use have begun [26–28]. However, all are limited to the 2D domain or thermal imaging. To address the need for well-annotated, dynamic 3D video of spontaneous facial behavior in response to meaningful and varied emotion inductions, we developed a *3D Dynamic Spontaneous* facial expression database with *annotation*, called the *Binghamton–Pittsburgh 4D Spontaneous Expression Database* (*BP4D-Spontaneous*), for the research community. The contributions of the work are as follows:

(1) We applied a series of well-designed tasks for authentic emotion induction. The tasks include social interviews between the previously unacquainted people (one a naïve subject and the other a professional actor/director), planned activities (e.g., games), film clip watching, a cold pressor test for pain elicitation, a social challenge to elicit anger followed by reparation, and olfactory stimulation to elicit disgust.
(2) Well-experienced, certified FACS coders annotated the videos. Frame-level ground-truth for facial actions was obtained.
(3) The effectiveness of the eliciting methods has been verified by subject self-report and FACS analysis.
(4) An alternative subjective evaluation and validation were conducted by human observer ratings.
(5) A set of meta-data, including AU codes, tracked 3D/2D features, and head poses is provided.
(6) Additionally, the quality and usefulness of the database have also been evaluated and validated through a number of applications in spontaneous facial expression recognition, 3D dynamic Action Units recognition, and a case-study for authentic pain expression analysis.

The remainder of this paper is organized as follows. In Section 2, we introduce the data acquisition procedure, expression elicitation method, and data processing and organization. In Section 3, we describe the creation of meta-data, including FACS coding, 3D/2D feature tracking, and head pose estimation. The experimental method and data quality are then evaluated through the analysis of the self-report information, subjective ratings, and AU statistics in Section 4. In Section 5, we further verify the usefulness of the database through applications in 4D spontaneous facial expression recognition and AU recognition, as well as a case study on pain expression analysis. Finally, concluding remarks and future work are given in Section 6.

2. High-resolution data acquisition

2.1. System setup

A Di3D dynamic face capturing system [12] captured and generated 3D facial expression sequences. The data include both 3D model sequences and 2D texture videos. The system consists of two stereo cameras and one texture video camera. The three cameras are placed on a tripod with two lighting lamps, one of each side of the cameras. A calibrating board and a blue board are used for calibration and background segmentation. With one master machine and three slave machines, the system captures the 3D videos at a speed of 25 fps. In addition to the 3D imaging system to capture the head-shoulder regions, we have also setup a regular video camera to capture the entire scene and audio for site monitoring and as a reference for possible audio-visual editing in the future if needed. The data is captured in the normal lighting conditions of an indoor lab environment. Fig. 1 shows an example of the imaging system at work.

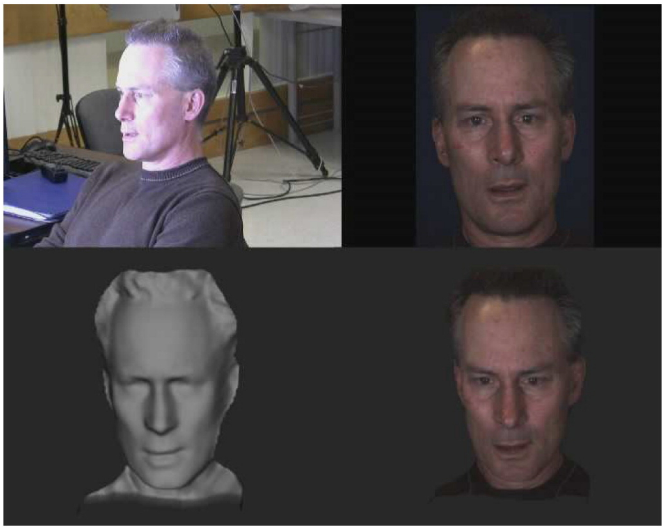


Fig. 1. Upper-left: general view from a regular camera; Upper-right: 2D video; Lower-left: 3D dynamic geometric model; Lower-right: 3D dynamic geometric model with mapped texture.

2.2. Data capture

Each participant was instructed to sit in front of the 3D face capturing system at about 51 inches distance from the cameras. After view adjustment and an initial preview capture, the capture procedure started by following an emotion elicitation protocol as described below.

2.2.1. Emotional expression elicitation

We define the "spontaneous facial expressions" as facial actions that are not deliberately posed, i.e., facial actions that occur in the course of social interaction or other social or non-social stimuli. For recording spontaneous affective behavior, a good trade-off between the acquisition of natural emotional expressions and data quality is needed. If the recording environment is too constrained, genuine emotion and social signaling become difficult to elicit. If the recording environment is unconstrained, substantial error may be introduced into the recordings. In the psychology literature, well-validated emotion techniques and guidelines have been proposed to meet this challenge [43].

To elicit target emotional expressions and conversational behavior, we used approaches adapted from other investigators plus techniques that proved promising in our pilot test. Each task was administered by an experimenter who was a professional actor/director of performing arts. The tasks include *interview*, *video-clip viewing and discussion*, *startle probe*, *improvisation*, *threat*, *cold pressor*, *insult*, and *smell*. Each task together with its target emotion is described in Table 1. Interviews elicit a wide range of emotion and interpersonal behavior [56,57,60]. Film clips and games [10,46,61] are well-validated approaches to elicit

Table 1
Eight tasks for emotional expression elicitation.

Task	Activity	Target emotion
1	*Interview*: talk to the experimenter and listen to a joke (interview).	Happiness or amusement
2	*Video clip*: watch a video clip and discuss it with the experimenter.	Sadness
3	*Startle probe*: sudden, unexpected burst of sound.	Surprise or startle
4	*Improvisation*: play a game in which the subjects improvise a silly song.	Embarrassment
5	*Threat*: anticipate and experience physical threat.	Fear or nervous
6	*Cold pressor*: submerge a hand in ice water for as long as possible.	Physical pain
7	*Insult*: experience harsh insults from the experimenter.	Anger or upset
8	*Smell*: experience an unpleasant smell.	Disgust

emotion. Cold pressor is well studied to safely elicit pain expressions without the risk of tissue injury [44]. Olfactory stimuli can reliably elicit disgust [62]. These methods evoke a range of authentic emotions in a laboratory environment [11].

After participants gave informed consent to the procedures and permissible uses of their data, the experimenter explained the general procedure (without giving any details about the specific tasks) and began the emotion inductions. Following usage in the psychology literature, each emotion induction is referred to as a "task". The experimenter was a professional actor and director. Each participant experienced eight tasks, as summarized in Table 1. Those tasks were seamlessly spaced with smooth transitions between them. Immediately after each task, participants completed self-report ratings of their feelings unless otherwise noted.

The protocol began with a conversation, which included joke telling, between the participant and the experimenter. The relaxed exchange and shared positive emotion were intended to build rapport and elicit expressions of amusement. After rating the first experience, the participant watched and listened to a documentary about a real emergency involving a child, followed by an interview that gave them opportunity to talk about their feelings in response to the task. Reactions of sadness were the intended responses.

Next, the participant was asked to participate in several activities with the experimenter. These included startle triggered by a siren, embarrassment elicited by having to improvise a silly song, fear while playing a game that occasioned physical danger, and physical pain elicited by submerging their hand in ice water. Following this cold pressor task, the experimenter intentionally berated the participant to elicit anger followed by reparation.

Finally, the participant was asked to smell an unpleasant odor to evoke strong feelings and expressions of disgust. The tasks concluded with a debriefing by the experimenter. Each task was recorded about 1–4 min and archived as described in sub-Section 2.3.

The procedures elicited a range of emotions and facial expressions that include happiness/amusement, sadness, surprise/startle, embarrassment, fear/nervous, physical pain, anger/upset, and disgust.

2.2.2. Participants

Forty-one participants (23 women, 18 men) were recruited from the departments of psychology and computer science as well as from the school of engineering. They were 18–29 years of age; 11 were Asian, 6 were African-American, 4 were Hispanic, and 20 were Euro-American (Table 3).

2.3. Data processing and database organization

For each task, there were three synchronized videos to be captured from two gray-scale stereo cameras and one color video camera. The stereo videos were processed by four machines (PCs) in parallel. Each pair of stereo images is processed using a passive stereo photogrammetry approach to produce its own range map. The range maps are then combined to produce a sequence of high-resolution 3D images. The geometric face model contains 30,000–50,000 vertices. The 2D texture videos are 1040 × 1392 pixels/frame. Fig. 2 shows example expressions of eight emotions elicited from eight tasks.

Table 2
The frame number for each task among 41 participants.

Task		#Frames
1	Interview	47,640
2	Video clip	65,555
3	Startle probe	12,863
4	Improvisation	60,647
5	Threat	52,323
6	Cold pressor	44,670
7	Insult	69,033
8	Smell	15,305

The database is structured by participant. Each participant is associated with eight tasks. For each task, there are both 3D and 2D videos. Table 2 illustrates the total number of frames for each task across all 41 participants in the database. Although tasks varied in duration, to reduce storage demands and processing time, each video consists of the segment during which the participant was most expressive (about one minute on average). This reduced the retention of frames in which little facial expression occurred. The video data are about 2.6 TB in size, and the average number of vertices for each 3D model is about 37,000.

Meta-data consists of manually annotated action units (FACS AUs), automatically tracked head pose, and 2D/3D facial landmarks. Table 3 summarizes the 3D dynamic spontaneous facial expression database. Fig. 3 shows the data structure of each task. Fig. 4 shows several samples of 3D dynamic spontaneous facial expression sequences. The meta-data (e.g., AU codes, tracked features, and head poses) will be described in detail in the next section.

3. Data annotation and meta-data creation

3.1. FACS coding

Automatic detection of FACS action units is a major thrust of the current research in automated facial image analysis [22]. To provide necessary ground truth in support of these efforts, we annotated facial expressions using the Facial Action Coding System (FACS) [17,18].

For each participant, we code action units associated with emotion and paralinguistic communication. Because FACS coding is time intensive, we prioritized coding to focus on 20-second segments that were most productive of facial expression.

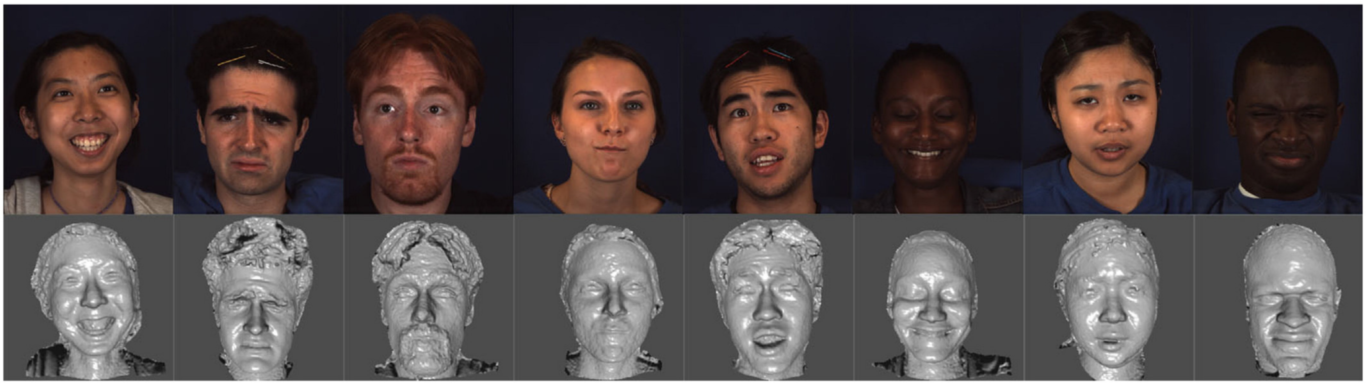


Fig. 2. 2D and 3D examples of eight emotional expressions from task 1 to task 8 (from left to right), respectively.

Table 3
Summary of BP4D-spontaneous database.

# of participants	# of tasks	# of 3D + 2D sequences	# of metadata sequences (i.e., annotated AUs, facial landmarks, and poses)
41	8	328	328

Note: Asian (11), African-American (6), Hispanic (4), and Euro-American (20).

For each of the eight tasks, FACS coders coded a 20-second segment that had the highest density of facial expression. Coders were free to code for longer than 20 s if the expression continued beyond that duration. If a video was less than 20 s, it was coded in its entirety. Descriptive statistics are reported in Table 4.

For each condition, two experienced FACS-certified coders independently coded onsets and offsets of 27 action units per the 2002 edition of FACS [36] using Observer Video-Pro Software [21]. These AUs, the corresponding amount of frames, and the number of AU events (from onset to offset) for each are listed in Table 5. An event is defined as a continuous series of AU from onset to offset. The observer system makes it possible to manually code digital video in stop-frame and at variable speed and later synchronize codes according to the digital time stamp. For AU 12 and AU 14, intensity was coded as well on a 0–5 ordinal scale using custom software.

To quantify the inter-observer exact (25 fps) agreement, the same thirty-six randomly selected video segments were coded by both coders. Their judgment was quantified using coefficient kappa [37], which is the proportion of agreement above what would be expected to occur by chance. Table 5 reports the kappa reliability.

In summary, the expression sequences were AU-coded by two experts. For each sequence, 27 AUs were considered for coding. For each of the target AUs, we have various numbers of coded events, where an event is defined as the contiguous frames from onset to offset.

3.2. 3D feature tracking

We defined 83 feature points around the 3D facial areas of eyes, nose, mouth, eyebrows, and chin contour in the initial frame of a video sequence (see Fig. 6(a)). Extended from the active appearance model approach [30], we applied our newly developed 3D geometric surface based temporal deformable shape model [40] to track 83 points on the 3D dynamic surface directly. Our developed method involves fitting a new multi-frame constrained 3D temporal deformable shape model (3D-TDSM) to range data sequences. We consider this a temporal based deformable model as we concatenate consecutive deformable shape models into a single model driven by the appearance of facial expressions. This allows us to simultaneously fit multiple models over a sequence of time with one 3D-TDSM.

To construct a temporal deformable shape model, we applied a representation of the point distribution model to describe the 3D shape, in which a parameterized model S was constructed by 83 landmark points on each model frame. Such a set of feature points (shape vector) was aligned by the Procrustes analysis method [30]. Principal component analysis (PCA) was then performed on the new aligned feature vector to retain 95% of the variance of the model. This was done to estimate the different variations of all the training shape data from the mean shape $\overline{S}$, as shown in Eq. (1). By adjusting the weight vector w in the range, the instances of 3D-TDSM could be constructed. When approximating a new shape S, the point distribution model was constrained by both the variations in shape and the shapes of neighbor frames. In our experiment, two neighbor frames were considered.

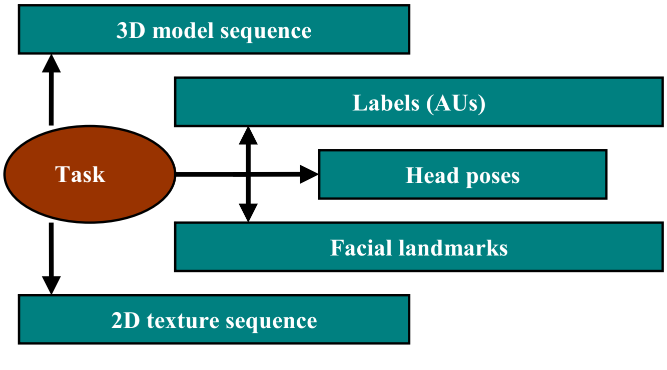


Fig. 3. Organization of each task in database.

$$S = \overline{S} + Vw \tag{1}$$

$$D = \sum_{i=1}^{N} \sqrt{(u_i - x_i)^2 + (v_i - y_i)^2 + (w_i - z_i)^2} \tag{2}$$

where N is 83. After creating the instance of 3D-TDSM, the 3D range model was fully searched to find the closest point to each landmark among the instances. Then, the weight vector w can be calculated. Since the shape variation was limited by both the deformation rules and the shapes of neighbor frames, if w was outside of the domain it was discarded. For the remaining acceptable candidates, the Procrustes distance, between the corresponding 3D-TDSM instance (u, v, w) and the result from fitting (x, y, z), was computed (as shown in Eq. (2)). The candidate with the minimum D value was chosen as the tracking result. Fig. 5 illustrates the fitting process, and Fig. 7 (lower row) shows several sample frames of the tracked 83 feature points on a 3D model sequence. The detailed algorithm is described in [40].

3.3. 2D feature tracking

Two-dimensional facial expression sequences were automatically tracked using the constrained local model (CLM) approach of [38,39]. Forty-nine landmark points were defined in the 2D face region (see Fig. 6(b)). All CLM tracking was reviewed offline for tracking errors. Coded were: 1) "Good tracking"; 2) "Multiple errors"; 3) "Jawline off"; 4) "Occlusion"; and 5) "Face out of frame". A confidence score of tracking performance was given for each frame. If the score is lower than the error threshold, the frame is reported as lost-track. Fig. 7 (upper row) shows several sample frames with the tracked points.

Note that the 3D-TDSM tracks the 83 feature points purely based on 3D geometry shape while the 2D-CLM tracks 49 feature points based on 2D images only. The lost-track rates of 3D-TDSM and 2D-CLM are 0.148% and 0.194%, respectively. The tracked 3D points offer features that are not reliably available from 2D tracking in case of large pose variations. Since the two sets of tracking points were obtained independently, they can be used for mutual verification and compensation, thus allowing researchers for further study of feature alignment between 2D and 3D, and developing algorithms for 2D/3D feature detection and tracking with comparison to the baseline 2D/3D features that we provide.

3.4. Head pose tracking

Head pose, which includes rigid head motion, is important for image registration and is itself of communicative value, (e.g., downward head pitch when coordinated with smiling communicates embarrassment.) Two sets of head pose data are provided based on 3D and 2D modalities. In 2D texture sequence, head pose was measured from the 2D videos using a cylindrical head tracker [19]. This tracker is person-independent, robust, and has concurrent validity with a person-specific 2D + 3D AAM [20] and with a magnetic motion capture device [19]. In 3D dynamic sequences, the 3D geometric features can derive the pose information directly using three points (i.e., two eye corners and one nose-base point). The head pose (pitch, yaw, and roll) was

Please cite this article as: X. Zhang, et al., BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database, Image Vis. Comput. (2014), http://dx.doi.org/10.1016/j.imavis.2014.06.002

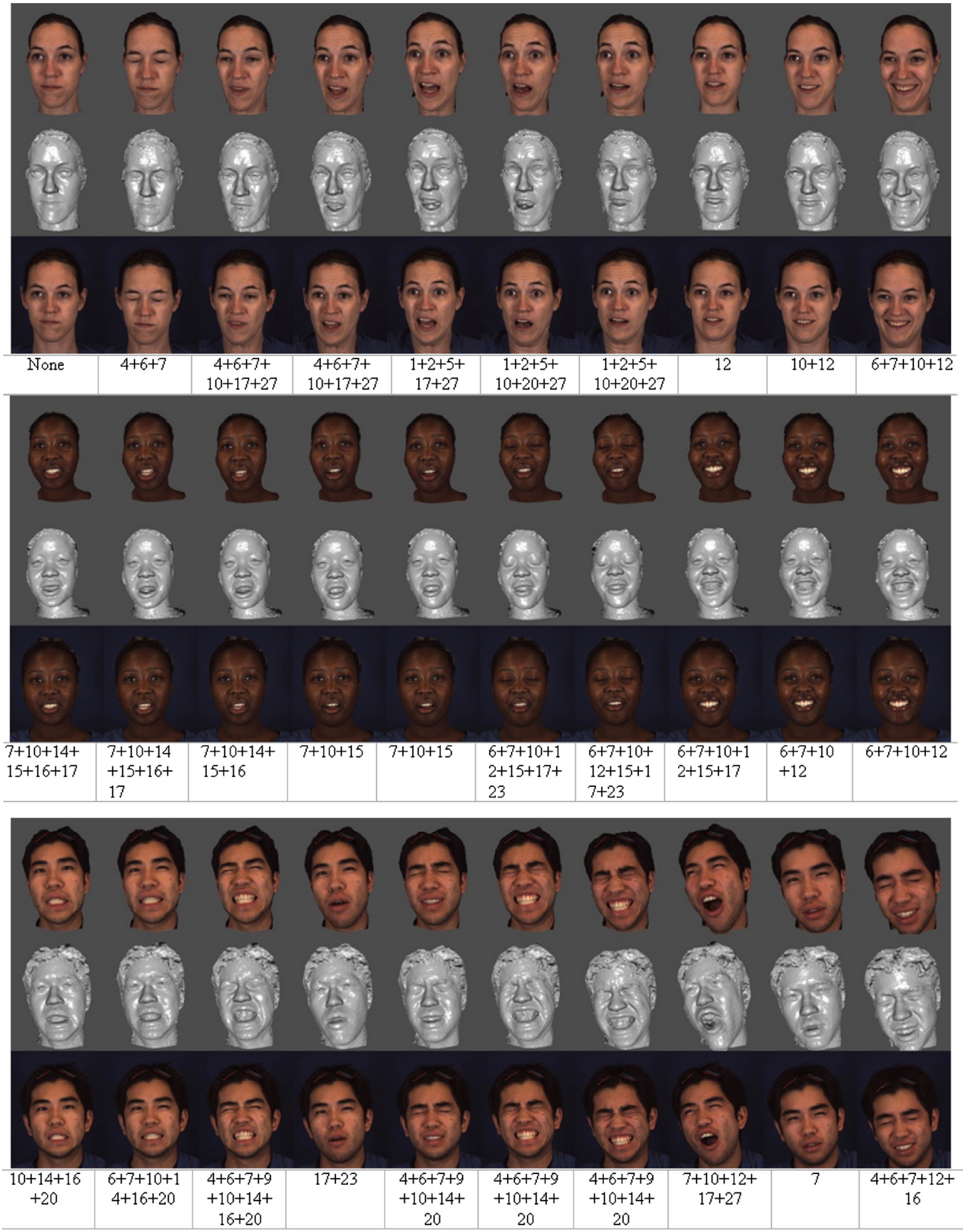


Fig. 4. Samples of textured models, shaded models, original 2D videos, and the annotated action units (AUs).

Table 4
Descriptive statistics for FACS-coded videos (unit of measure is seconds).

Task	Activity	*Minimum*	*Maximum*	*Mean*
1	*Interview*	13.00	29.71	19.67
2	*Video clip*	12.12	25.00	20.21
3	*Startle probe*	8.56	16.76	12.25
4	*Improvisation*	16.14	24.12	19.74
5	*Threat*	18.53	31.00	20.74
6	*Cold pressor*	8.00	23.00	18.95
7	*Insult*	17.24	25.01	19.91
8	*Smell*	3.60	21.40	11.49

Note. Unit of measure is seconds. Data are based on video from all 41 participants.

measured with respect to the frontal pose. Table 6 shows the proportion of frames which differs from the frontal view. In the case of extreme head movement, which causes over 50% partial occlusion of facial regions, the pose information is not obtainable; thus, the frame is labeled as lost-track.

Since the head pose information has been derived from two different modalities with independent algorithms, the lost-track errors occur only when both pose tracking methods fail. As a result, the lost-track error of pose is only 0.063%. The error is reduced as compared to the individual error rates of the two tracking methods (as indicated in Section 3.3).

In short, the 3D-TDSM and the 2D-CLM provide the tracking results of BP4D-Spontaneous in two modalities. The accuracy of head pose tracking is increased due to the combination of the two tracking results.

4. Evaluation and analysis

In order to evaluate the effectiveness of the emotion elicitation, we analyze the data statistically based on participants' self-report, subjective ratings from naïve observers, and AU distributions from coded videos.

Table 5
Descriptive statistics for kappa reliability, events, and frames.

Action unit	Name	Kappa reliability	Events	Frames
1	Inner brow raiser	0.90	474	31,043
2	Outer brow raiser	0.95	380	25,110
4	Brow lowerer	0.92	408	29,755
5	Upper lid raiser	0.97	182	5693
6	Cheek raiser	0.91	540	67,677
7	Lid tightener	0.92	569	80,617
9	Nose wrinkler	0.91	140	8512
10	Upper lip raiser	0.90	591	87,271
11	Nasolabial deepener	0.94	33	7184
12	Lip corner puller	0.92	448	82,531
13	Cheek puller	n/a	2	138
14	Dimpler	0.92	571	68,376
15	Lip corner depressor	0.79	657	24,869
16	Lower lip depressor	0.68	219	6593
17	Chin raiser	0.88	1203	50,407
18	Lip pucker	0.83	33	568
19	Tongue show	0.84	61	1197
20	Lip stretcher	0.95	105	3644
22	Lip funneler	0.95	46	606
23	Lip tightener	0.78	805	24,288
24	Lip pressor	0.86	457	22,229
27	Mouth stretch	0.95	55	1271
28	Lip suck	0.97	117	5697
30	Jaw sideways	0.95	15	506
32	Bite	0.98	26	1466
38	Nostril dilator	0.94	1	1319
39	Nostril compressor	0.97	25	657
Overall		0.90	8161	639,224

Note: Data are based on video from all 41 participants. Overall kappa is weighted average based on 36 double-coded videos. An event is defined as a set of contiguous frames from onset frame to offset frame.

4.1. Participants' self-report & analysis

After each task, participants used 6-point Likert-type scales (0 to 5, none to extremely) to report their felt emotions for each task. The emotions listed were *relaxed*, *happiness/amusement*, *disgust*, *anger/upset*, *sadness*, *sympathy*, *surprise*, *fear/nervous*, *embarrassment*, *physical pain*, *and startle*. As has been found previously [53], participants could and did experience more than one emotion for each task. Fig. 8 shows the highest rated emotions reported for each task. Similarly, Table 7 compares the most highly-rated emotion and the target emotion for each task. Except for task 7, the target emotion for each task (see Table 1) was the one most highly rated by the majority of participants. For instance, the highest bar of task 8 shows that the majority of subjects rated the "disgust" emotion as the main emotion for that task. The highest bar of task 6 shows the majority of subjects rated the "pain" feeling as the main emotion. Accordingly, almost all of the other tasks show this property as well. For task 7, the most highly rated emotions could all be expected from the context (anger at the experimenter, embarrassment from not doing better). Overall, the tasks generally succeeded in evoking the target emotions.

To study the emotion distribution, we analyze all scales except 0 (None) of the self-report rating of each task and illustrate the emotion scale distribution in Fig. 9. The markers on each line show the intensity percentage among the votes for the corresponding emotion category. Note that one to three emotions for each task are selected for illustration. Our selection criteria are based on the results in Fig. 8; for each task, we select a given emotion for illustration if its vote count in Fig. 8 makes up over 20% (approximately 13 votes) of the total votes for the task. In general, Fig. 9 supports the findings of Fig. 8 in that the major emotions of each task illustrated in Fig. 8 have high grades (scale 3 and above) in Fig. 9. In other words, the high grades (from scale 3 to the highest scale 5) account for the majority of the votes from the 41 subjects for the target emotions of each task.

Most of the tasks could elicit multiple emotions or mixed emotions. However, there is still a principal emotion for each task. For the tasks such as task 2 (Documentary for Sadness), task 3 (Burst of Sound for Surprise/Startle), task 6 (Cold Pressor for Physical Pain), and task 8 (Smell for Disgust), the majority of votes are distributed in the highest (or second highest) grades, showing that the target emotion elicitation of those tasks was successful and the intensity of the corresponding emotions is strong. For task 7 (Insult for Anger/Upset), although the intensity of "anger/upset" is not very strong, the "anger/upset" emotion is still a major emotion of this task.

In general, the self-report information shows the target emotions were elicited effectively.

4.2. Subjective rating and analysis

4.2.1. Expression labeling from naïve observers

To evaluate the results of the emotion elicitation, we conducted a subjective rating experiment by asking naïve observers to label each video with two of the eight expressions for all the tasks.

Five naïve observers were recruited to participate in the rating experiment. First, the purpose of the experiment was explained to the observer. Then, the observer watched the AU coded video segments of all subjects in the database. The videos were presented randomly in order to avoid any ordering effect, and the videos were also muted to ensure that their rating was only based on the visual information. For each video segment, eight expression labels were provided to choose from. The observer was asked to choose up to two most likely expressions in the order of their confidence. To indicate the confidence level of each choice, the observer used a three-scale list which represents *high confidence*, *moderate confidence*, and *low confidence*. The observer was allowed to replay the video segment if needed. A timer was started when a given video segment was started, and it was stopped when the final choices for the segment were made. In this way, the expression

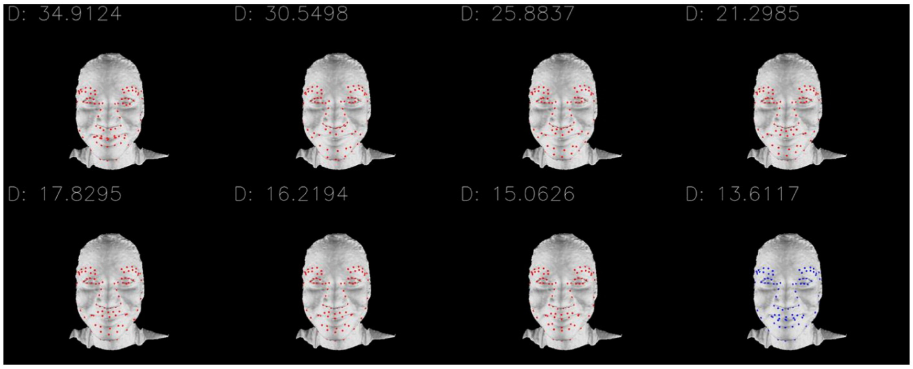


Fig. 5. Eight samples of the 3D-TDSM candidates. The distance score is listed at the top of each candidate. The one with the minimum score (bottom right) is the best fit.

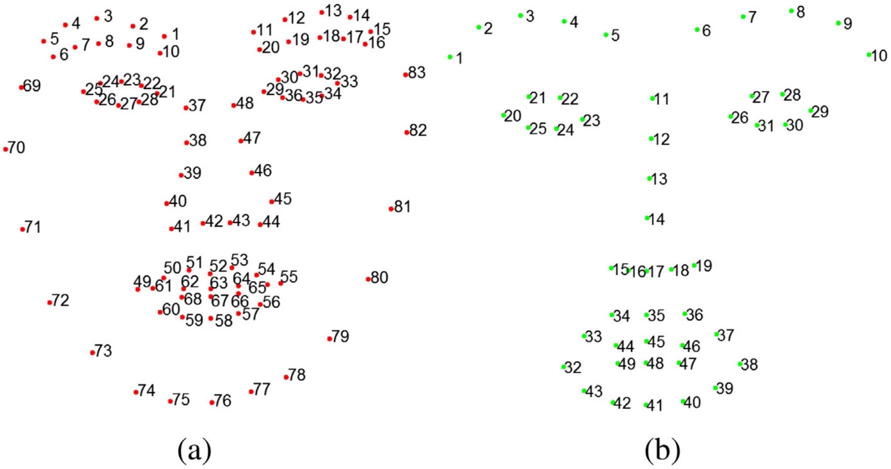


Fig. 6. (a) 3D and (b) 2D landmarks' indices.

rating and judgment time were also recorded. If the target emotion of a task was recognized by an observer correctly, which means that the corresponding target expression was included in the top two choices by the observer, we count it as a correct recognition. The results show that the correct recognition rates of five naïve observers are 61% (happiness), 77.1% (sadness), 94.6% (startle), 65.9% (embarrassment), 73.2% (fear), 66.8% (pain), 55.1% (anger), and 83.4% (disgust).

4.2.2. Analysis of subjective ratings

The inter-rater reliability was examined using Fleiss' kappa coefficient [37]. This has been used to assess the reliability of agreement between raters when assigning categorical ratings to a number of items. In our case, five raters assigned eight expression categories to all 328 video segments. The kappa value is in the range of −1 to 1, corresponding to a negative range (−1, 0) and a non-negative range (0, 1). The

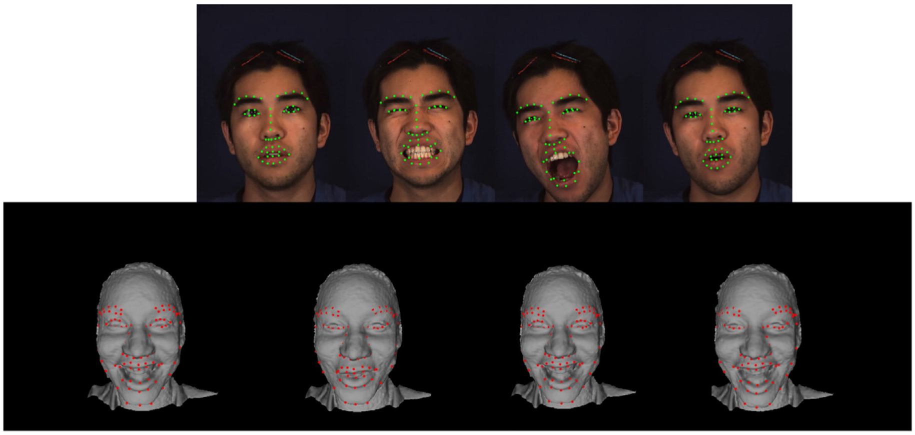


Fig. 7. CLM-tracked feature points on a 2D sequence of a male subject (upper row); a sample 3D sequence with 3D-TDSM tracked feature points of a female subject (lower row).

Table 6
Proportion of frames in different pose relative to the frontal view.

	Pitch	Yaw	Roll
<5°	60.3%	80.0%	86.4%
<10°	94.7%	97.5%	98.0%
<15°	99.5%	99.6%	99.8%
<20°	99.9%	99.8%	99.9%
>20°	0.1%	0.2%	0.1%

Table 7
Major emotions elicited from each task based on self-report.

Task	Target emotion	Emotion most reported
1	Happiness or amusement	Happiness/amusement
2	Sadness	Sadness
3	Startle or surprise	Startle, surprise
4	Embarrassment	Embarrassment
5	Fear or nervous	Nervous/fear
6	Physical pain	Physical pain
7	Anger or upset	Anger/upset, embarrassment
8	Disgust	Disgust

non-negative value between 0 and 1 can be divided into 0.2 sized steps. Therefore, a total of 6 levels can be generated. Landis and Koch interpret the six levels as *poor*, *slight*, *fair*, *moderate*, *substantial*, and *almost perfect* agreement, respectively [63].

Table 8 shows the Fleiss' kappa value for 8 expression categories. The overall kappa indicates moderate agreement between all observers.

To further evaluate the performance, observers' confidence level and their judgment time factor are also studied. We assigned to the confidence levels *low*, *moderate*, and *high* the values *0*, *1*, and *2*, respectively. The judgment time factor is a value of a given judgment time divided by the length of the corresponding video segment of a task. If an observer spends more time to make a decision than the length of the video segment, the factor of the task is larger than 1. Otherwise, the factor is less than 1 if less time is used. Table 9 lists the average values of confidence levels and the average values of judgment time factor.

To assess the performance of the human observers, a confusion matrix with respect to the accuracy of the expression ratings from the five naïve observers is illustrated in Table 10. The diagonal line shows the correct recognition rates. Among them, *startle*, *disgust*, and *sadness* are among the three most distinguishable spontaneous expressions with the three highest recognition rates by observers. *Happiness* is sometimes confused with *embarrassment* as people sometimes show a soft smile in an embarrassing situation. For the improvisation task, subjects felt embarrassed, and, in the insult task, subjects could also feel embarrassed if they thought they had not performed well. It is not unusual for people to restrain their genuine negative emotions in social activities. People are likely to show a soft smile in an awkward situation to defuse tension. Thus, the smile could be misread by observers.

The diversity and variety of spontaneous expressions still pose a challenge for human observers to distinguish them. However, in general, the correct classification rates for all the expressions in Table 10 are still dominant when compared to the misclassification rates. The results are comparable to the results obtained with machine recognition (which will be described in Section 5.1). This shows that our elicitation method is effective when eliciting distinctive facial activities associated with different tasks. The distinctive dynamic information exhibited in spontaneous expressions benefits expression reading by human observers.

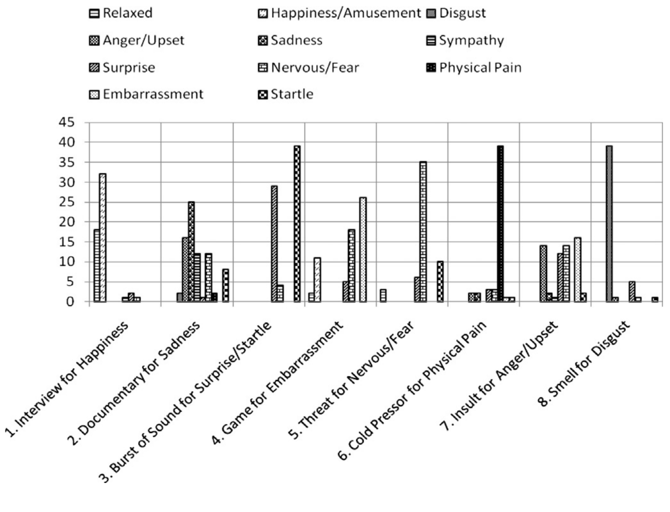


Fig. 8. Statistics of self-report emotion distribution for task 1 to task 8 (from left to right); vertical axis is the number of votes.

4.3. Action unit analysis

4.3.1. AU distribution in spontaneous expressions

Holistic expressions of emotion can be defined in terms of one or more anatomic actions [52,54,55]. Following the previous work [16, 48–50], Table 11 shows a mapping between AU(s) and the hypothetical emotion. Table 12 shows the 27 action unit distributions across the eight tasks. The value is a percentage *P* that is defined as follows:

$$P = N_{\mathrm{AUi}}/N_{\mathrm{t}} \qquad (3)$$

where N_{AUi} is the number of frames that show the *i*th coded AU of a task (i = 1, 2, …, 27), and N_{t} is the total number of AU coded frames of the task (t = 1, 2, …, 8). In this section, we report the extent to which holistic expressions defined using AUs corresponded to the target emotions.

The AU distribution among different elicitation tasks illustrates the complexity and diversity of spontaneous expressions. As we can see in Table 12, 41 subjects show genuine expressions which cover all the 27 action units. Also, the frequencies of different AU occurrence vary dramatically across both AUs and tasks. Unlike posed expressions, the facial action of spontaneous expressions appears involuntarily. In Table 11, AU 12 is related to happiness, and indeed a standard happy face contains AU 6 + 12. In Table 12, it shows that column T1 has a major occurrence of AU 6 (61.7%) and AU 12 (81.9%). However, it does not mean that AU 12 is unique to happiness. In fact, AU 12 occurred with varying frequencies in all tasks. This is consistent with the hypothesis that AU 12 (smiling) is not specific to positive emotion. While AU 12 is a defining feature of expressed enjoyment, it occurs in pain, embarrassment, and other negative emotions as well as in depression [51,58,59]. In Table 11, we know that *disgust* has either AU 9 or AU 10 present, and task 8 (experience an unpleasant smell) gives the highest percentages of AU 9 (27.8%) and AU 10 (72.1%) among all tasks. AU 4 and AU 9 are related to negative emotions. AU 4 shows relatively high percentages in negative emotion tasks such as T2 (43.6%), T6 (31.0%), and T8 (43.3%), while it shows low percentages in tasks for happiness (7.1%). Similar evidence can be found for AU 9.

4.3.2. Effectiveness of elicitation

Besides participants' self-report and observers' rating report, we further study the AU distribution to verify the effectiveness of our emotion elicitation method objectively.

Using the criteria listed in Table 11, we examine whether some AUs of the target emotions appear most frequently in the corresponding tasks. According to the percentage of AU sets that each task has in Table 12, the top two ranked tasks (in order from left to right) are illustrated in Table 13. The first column lists the AU criteria to be used for matching the corresponding target emotions (as the second column shows). For example, '12, 24' means that either AU 12 or AU 24 needs to appear for the target emotion *Embarrassment*; and '1 + 2 + 4, 1 + 2 + 5' means

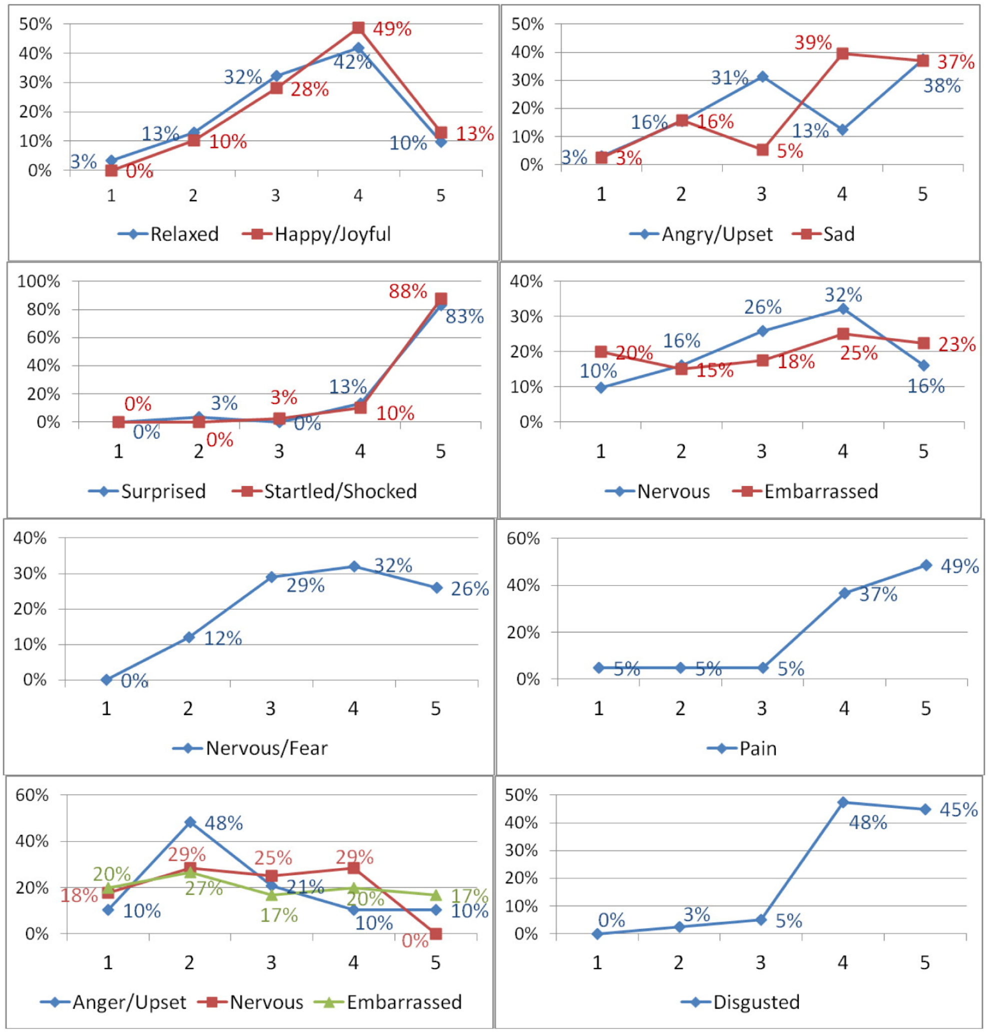


Fig. 9. Self-reported emotion distribution across 5 scales (very slightly, a little, moderately, quite a bit, extremely) in each of the eight tasks (from left to right, top to bottom, the charts correspond to the tasks 1–8 as shown in Fig. 8).

either AU 1 + 2 + 4 or AU 1 + 2 + 5 needs to appear for the target emotion *Fear*/*Nervous*. The third column shows the top two ranked tasks based on the corresponding AU counts from the spontaneous expressions of 41 subjects. It shows that the corresponding AUs appeared most frequently in the top two tasks. Moreover, these two top tasks always include the expected target emotion for all the cases (bold font). This verifies that our elicitation method is effective in eliciting the target emotions. As an example, in the second row, AU 4 is used as a criterion for matching the *sadness* (task 2). As a result, AU 4 has the most frequent occurrence in task 2 based on Table 12. Therefore, the top ranked task is the *sadness* task, which is exactly the expected target emotion.

Therefore, we can see that the AU distribution of the BP4D-Spontaneous database verifies the distinctiveness of the spontaneous expressions. Similar to the results from the self-reported emotions, the findings for holistic expressions suggest that the tasks were effective in eliciting the target emotions.

5. Validation & application in facial expression analysis

5.1. 4D spontaneous facial expression recognition

To validate the data for spontaneous facial expression recognition, we apply an existing 3D dynamic facial expression descriptor [42] for

Table 8
Kappa coefficients from multiple raters for expression categories.

Expression	Kappa value
Happiness/amusement	0.4325
Sadness	0.6335
Startle	0.8366
Embarrassment	0.4128
Fear	0.4933
Physical pain	0.5250
Anger	0.3724
Disgust	0.7193
Overall	0.5535

Table 9
Average value of confidence level and judgment time factor for expression categories.

Expression	Conf. level	Judgment time factor
Happiness/amusement	1.810	1.059
Sadness	1.780	1.139
Startle	1.941	0.962
Embarrassment	1.868	1.039
Fear	1.932	0.910
Physical pain	1.839	1.041
Anger	1.824	1.159
Disgust	1.941	1.162

Table 10
Confusion matrix for the relationship between target expression and observers' recognition.

Rec. / Tar.	Hap.	Sad.	Star.	Emb.	Fear	Pain	Ang.	Dis.
Hap.	0.610	0.010	0.029	0.166	0.029	0.020	0.112	0.024
Sad.	0	0.771	0.029	0	0.093	0.029	0.044	0.034
Star.	0	0.020	0.946	0	0.010	0.020	0	0.005
Emb.	0.249	0	0.010	0.659	0.034	0.005	0.029	0.015
Fear	0.049	0.005	0.039	0.059	0.732	0.044	0.059	0.015
Pain	0.024	0.059	0	0.044	0.068	0.668	0.093	0.044
Ang.	0.161	0.039	0.015	0.122	0.078	0.024	0.551	0.010
Dis.	0.015	0.015	0	0.015	0.049	0.054	0.020	0.834

expression classification. A Hidden Markov Model (HMM) is used to learn the temporal dynamics and spatial relationships of facial regions. To do so, a generic model is adapted to each range model of a 3D model sequence. The adaptation is controlled by a set of 83 key points based on the radial basis function (RBF). After adaptation, the correspondence of the points across the 3D range model sequence is established. We apply a surface labeling approach [42] to assign each vertex one of eight primitive shape types. Thus, each range model in the sequence is represented by a "label map". We use Linear Discriminative Analysis (LDA) to transform the label map to an optimal compact space to better separate different expressions. Given the optimized features, an HMM is trained for each expression. Note that the HMM is applied to the optimal features of the label map rather than the trajectories of 83 landmarks. In recognition, the temporal/spatial dynamics of a test video is analyzed by the trained HMMs. As a result, the probability scores of the test video to each HMM are evaluated by the Bayesian decision rule to determine the expression type of the test sequence.

We conducted a person-independent experiment on 41 subjects. Following a 10-fold cross-validation procedure, we used 39 subjects for training and 2 subjects for testing, and achieved an average correct recognition rate of 73.7% for distinguishing eight spontaneous emotional expressions.

In order to make a comparison to the existing work [15,42], where *six prototypic posed* 3D dynamic facial expressions were used for recognition, we have also conducted an experiment for *six prototypic spontaneous* 3D dynamic expression recognition using the same 41 subjects. The results show that the correct recognition rate is 76.1%. Note that spontaneous expressions are more difficult to classify than posed expressions. When the same approach was applied to the 3D posed dynamic facial expression database BU-4DFE [15], a recognition rate for classifying six posed prototypic expressions is 83%. The performance degradation on classifying 3D spontaneous expressions is due to the complexity, mixture, and subtlety of the spontaneous expressions in the new database.

To further evaluate our approach, we conducted a comparison study by implementing the 3D static model-based approach using geometric primitive features [29] and the 2D texture-based approach using Gabor-wavelet features [31] to classify eight expressions from the entire BP4D-Spontaneous database. Note that we used two approaches in the case of the static image or static model. In the first approach, we chose an apex frame of each video sequence for the experiment. In the second approach, we chose three frames (a frame between onset and apex, apex frame, and frame between apex and offset); we then applied expression classification on the three frames individually. The output confidence levels (output scores measured by the probability of Naïve Bayesian Classifier) for the three frames were fused by averaging the output scores for each expression. Among 8 expressions, we chose the highest fused score as the recognized expression. The average recognition rates for the static model and static frame based approaches were 62.4% and 63.2%, respectively. The average recognition rates for the three-models and three-frames-based approaches were 65.7% and 66.8%, respectively. The three-frame fusion approach does show improved performance; however, the results are not as good as those from the 3D dynamic model-based approach (i.e., 73.7% as shown above for distinguishing eight spontaneous expressions from the entire database).

Table 11
Emotion description in terms of facial action units.

Target emotion	Criteria
Happiness or amusement	AU 12 present
Sadness	Either AU 1 + 4 + 15 or 11 or AU 6 + 15
Surprise or startle	Either AU 1 + 2 or 5 must be present for surprise AU 7 for startle
Embarrassment	AU 12 or 24
Fear or nervous	AU 1 + 2 + 4 or AU 1 + 2 + 5
Physical pain	AU 4, 6, 7, 9, 10
Anger or upset	AU 23 and 24 must be present in the AU combination
Disgust	Either AU 9 or 10 must be present

Table 12
27 action units percentage (%) in all tasks.

Task / AU	T1	T2	T3	T4	T5	T6	T7	T8
1	13.6	21.7	18.1	21.2	38.9	14.8	18.1	20.6
2	15.7	8.9	18.0	23.6	27.9	10.7	17.5	13.0
4	7.1	43.6	19.3	9.0	11.2	31.0	6.3	43.3
5	2.4	7.0	7.9	2.6	4.5	1.0	4.7	0.9
6	61.7	6.5	26.2	75.6	60.0	39.9	44.7	48.4
7	62.2	28.4	34.7	69.1	65.1	57.1	52.8	68.4
9	1.3	0.1	4.4	2.2	7.2	10.4	1.5	27.8
10	71.1	16.1	28.8	81.5	76.1	57.2	67.1	72.1
11	9.1	0.1	4.1	3.9	4.8	7.2	3.7	7.3
12	81.9	2.3	32.5	89.7	79.3	36.8	67.3	48.8
13	0	0	0	0	0	0.7	0	0
14	50.6	21.1	36.5	56.2	54.6	49.6	53.2	48.5
15	16.4	6.4	8.6	16.5	22.9	10.9	22.4	35.7
16	4.3	2.2	4.3	5.1	6.0	6.1	5.0	1.9
17	32.9	30.8	21.3	35.0	37.5	36.1	32.8	49.6
18	0.4	0.5	0	0.6	0.7	0.1	0.3	0.4
19	0.5	0.5	1.1	1.0	1.3	0.9	0.7	0.5
20	0.8	0.2	2.7	4.0	1.1	6.2	1.4	4.5
22	0.4	0	0.2	1.7	0.1	0.1	0.6	0
23	14.7	8.5	9.7	19.3	17.5	24.5	19.3	16.9
24	11.8	15.8	11.5	10.1	15.4	26.1	10.3	21.7
27	0.9	1.2	0.9	0.5	0.3	1.1	1.6	0.2
28	2.6	6.7	2.5	1.5	5.4	7.5	2.2	1.0
30	0.1	0.3	0.6	0.3	0	1.4	0.1	0.1
32	1.8	0	0.3	1.0	2.5	1.2	0.5	0
38	0.3	1.1	1.5	0	0.2	3.3	0.6	0.3
39	0.1	0.8	0.1	0.3	0.2	1.4	0	0.6

5.2. Cross-database 4D facial expression classification

We also performed a cross-database validation on our new 4D spontaneous database. A posed facial expression database (BU-4DFE [15]) was used for training, and the spontaneous database was used for testing.

Table 13
Top two tasks based on FACS emotion description[a].

AU criteria	Target emotion	Top two tasks indices
12	Happiness/amusement	Embarrassment, **happiness/ amusement**
4	Sadness	**Sadness**, disgust
5	Surprise/startle	**Startle**, sadness
12, 24	Embarrassment	**Embarrassment**, happiness/ amusement
1 + 2 + 4, 1 + 2 + 5	Fear/nervous	**Fear/nervous**, startle
4 + 6 + 7 + 9 + 10	Physical pain	Disgust, **physical pain**
23	Anger/upset	Physical pain, **anger/upset**
9 + 10	Disgust	**Disgust**, physical pain

[a] In the third column, the corresponding target emotion is in bold font.

Table 14
Spontaneous action unit recognition accuracy results: "All" = all blocks used; "Best" = best blocks used.

AU	Nebula		LBP-TOP 2D		LBP-TOP depth	
	All	Best	All	Best	All	Best
1	54.1%	58.4%	57.9%	49.4%	52.4%	48.5%
2	63.0%	64.8%	59.2%	55.4%	55.9%	53.1%
4	58.7%	63.1%	53.3%	48.4%	51.1%	48.9%
6	67.6%	68.8%	64.8%	64.8%	61.3%	61.7%
7	58.9%	58.0%	55.4%	51.9%	52.4%	53.2%
10	66.4%	65.9%	62.1%	54.3%	56.9%	58.6%
12	57.3%	57.8%	59.1%	54.7%	53.3%	54.7%
14	54.5%	59.1%	52.3%	46.4%	52.8%	48.1%
15	66.0%	69.0%	64.5%	61.6%	63.1%	62.1%
17	61.8%	65.6%	60.0%	44.6%	53.3%	43.5%
23	60.6%	61.4%	58.5%	53.4%	59.3%	55.5%
24	67.1%	67.6%	63.4%	56.3%	62.9%	54.0%
Avg.	61.3%	63.3%	59.2%	53.4%	56.2%	53.5%

We extended the idea of a 3D surface primitive feature into 4D space and developed a new feature representation: the so-called "Nebula" features [41]. Given each vertex on the face, a local spatiotemporal volume is built from the neighbor points across all frames in a time window (in our experiment, window size is 15 frames). Spatial neighborhood radius sizes of 3, 5, and 7 millimeters were tested, and the corresponding spatial voxel dimensions were 7 × 7, 11 × 11, and 15 × 15, respectively. The neighborhood data are voxelized (with x, y, t as the dimensions and depth z as the values) and fit to a cubic polynomial:

$$\begin{aligned} f(x;y;t) = {} & A\frac{1}{2}x^2 + B\frac{1}{2}y^2 + C\frac{1}{2}t^2 + Dxy + Ext + Fyt \\ & + Gx^3 + Hy^3 + It^3 + Jx^2y + Kx^2t + Lxy^2 + My^2t + Nxt^2 \\ & + Pyt^2 + Qx + Ry + St + U = z. \end{aligned} \tag{4}$$

The principal curvature directions and values are computed from the eigenvectors/eigenvalues of the Weingarten matrix:

$$\begin{pmatrix} A & D & E \\ D & B & F \\ E & F & C \end{pmatrix}. \tag{5}$$

A label is assigned based on the principal curvature values and a threshold T. Although this label compactly describes the shape of the feature, it does not give us its orientation. This orientation information is vital, since it can give us an indication of whether the surface changes across time or not. For example, consider two features with cylindrical shapes. Both have the same label; however, one may be oriented along the time axis, indicating no change across time, while the other may be perpendicular to the time axis, indicating a sharp change across time. Therefore, we use the label as well as the polar angles of the direction of least curvature as the values for each feature. The face area is then divided into regions. A 3D histogram is built for each region of the face with the features in that region. The variables for that histogram are the shape label and the two polar angles for each feature. The concatenated histograms from each of the regions give us our final feature vector. The construction of a single Nebula feature is shown in Fig. 10.

We selected data (happiness, disgust, and neutral) from BU-4DFE for training. We then tested directly on the spontaneous data using the Nebula feature approach. Our overall average accuracy was 71%. From the experiment, we find that Happy-Onset is often mistaken for Disgust-Onset. One possible explanation is that the spontaneous smiles, since they are mostly genuine smiles, also show activity around the eyes similar to some of the Disgust expressions; the posed smiles from BU-4DFE do not always demonstrate this. To the best of our knowledge, this is the first time that a cross-database test on 4D expression data has been performed. Spontaneous expression data is almost invariably more difficult to classify than posed expression data. As a consequence, we believe our results are encouraging.

5.3. Action unit recognition on spontaneous 4D data

We also performed experiments in AU recognition on BP4D-Spontaneous. We selected 16 subjects and tested on 12 AUs using a support vector machine classifier. Please note that we only tested on the AUs listed in Tables 14 and 15. In the database, the AUs are marked as present or not present for each annotated frame. Several 9-frame windows were extracted around several transition points (either near the apex frame or near the offset frame) for each AU. Segments without the AU in question were also extracted. With three different states per AU ("AU Onset", "AU Offset", and "No AU"), tests were conducted on each AU individually.

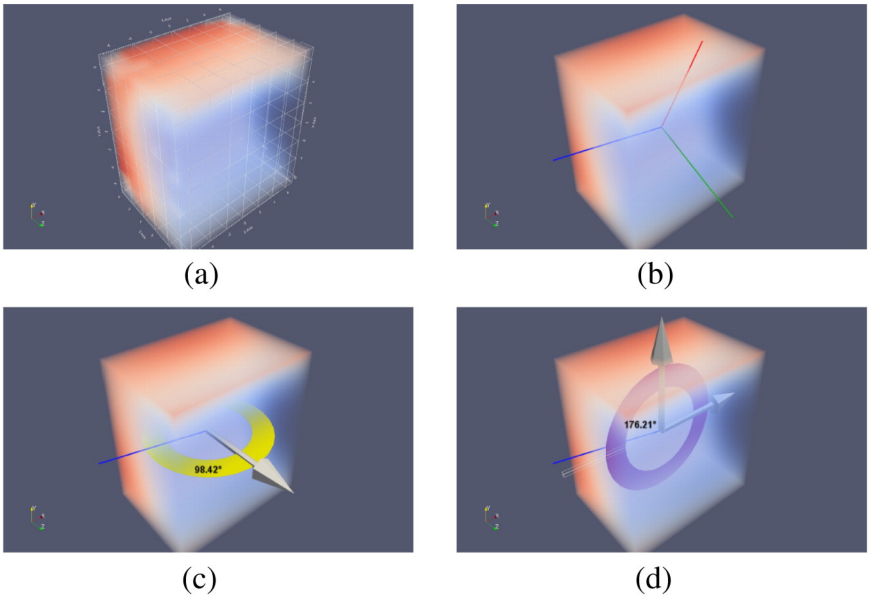


Fig. 10. 4D Nebula feature construction: (a) sample voxel; (b) resulting polynomial volume; principal axes shown in red, green, and blue, in order of curvature magnitude (label = 14, cylinder); (c) φ (phi) angle of least curvature direction from the time axis (shown as Z axis in image); (d) θ (theta) angle of least curvature direction in the XY plane. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 15
Spontaneous action unit recognition AUC score results: "All" = all blocks used; "Best" = best blocks used.

AU	Nebula		LBP-TOP 2D		LBP-TOP Depth	
	All	Best	All	Best	All	Best
1	0.621	0.640	0.631	0.584	0.627	0.574
2	0.682	0.697	0.642	0.586	0.614	0.602
4	0.657	0.700	0.637	0.578	0.589	0.601
6	0.796	0.800	0.774	0.783	0.768	0.758
7	0.696	0.680	0.678	0.662	0.660	0.648
10	0.791	0.773	0.739	0.668	0.686	0.705
12	0.703	0.726	0.732	0.691	0.653	0.663
14	0.682	0.717	0.648	0.605	0.624	0.583
15	0.739	0.784	0.688	0.626	0.731	0.705
17	0.758	0.801	0.713	0.607	0.681	0.598
23	0.704	0.722	0.675	0.614	0.702	0.639
24	0.774	0.791	0.713	0.633	0.720	0.592
Avg.	0.717	0.736	0.689	0.636	0.671	0.639

Table 16
P score of paired *T*-test on discrimination between pain and other emotion tasks on AU 6, AU 9, and AU 10.

Task	T1	T2	T3	T4	T5	T7	T8
T6	0.01	4.50×10^{-7}	3.52×10^{-4}	4.15×10^{-5}	3.08×10^{-4}	0.03	0.02

Various parameter combinations and regular grid block region configurations were tested. Specifically, block configurations with 15, 24, 35, 54, 77, and 96 blocks were tested. The number of rows and columns in each block region configuration was chosen to make the individual blocks square, given the size of the facial region. In addition to using all of the blocks of an entire face for a given configuration, we also tested using the "best" blocks for each AU. The best set of blocks is chosen based on where the AU appears on the face. The best blocks are in effect either dividing the face into top and bottom halves or taking the middle third of the face. For example, since AU 15 occurs around the mouth, the "best" region blocks for that AU are all the blocks in the lower half of the face; that is, the blocks on the top part of the face are excluded. In cases where the division of blocks was uneven, blocks were included rather than excluded (for instance, if the face was divided into 5 rows and we needed the top half of the face, rows 1, 2, and 3 were chosen).

The Nebula feature approach was again employed here. We also tested using the LBP-TOP (Local Binary Patterns from Three Orthogonal Planes) [47] approach on the re-rendered, pose-normalized texture images and the corresponding depth images. The accuracy results for all three approaches are shown in Table 14, while the AUC (Area Under Receiver Operating Characteristic Curve) score results are shown in Table 15. The confusion tables using the Nebula feature approach for each AU (using the parameters yielding the best accuracies) are shown in Table 17.

The average recognition AUC score using the ideal Nebula approach for each AU (i.e., all or "best" blocks) was over 0.738. This validates the usefulness of the new database as well as demonstrates the effectiveness of the test approaches. The highest average accuracy among all approaches is 63.3% from Nebula on the "best" blocks. Thus, we achieve over 4% better accuracy on average with the Nebula approach over the most accurate LBP-TOP approach (LBP-TOP 2D on all blocks). The AUC scores follow the same pattern, with Nebula on the "best" blocks giving us 0.736 AUC while the best LBP-TOP score (LBP-TOP 2D on all blocks) is 0.689. In Table 17, a common source of confusion is the misclassification between onsets/offsets and non-existence of an AU. This may due to the fact that spontaneous expression data is more challenging for AU recognition than that for posed expression data.

Note that in Tables 14 and 15, if we compare the "best" region results between Nebula and LBP-TOP 2D for individual AUs, the former is always better than the latter one. For the "all" region test, this is also true except for AU 1 and AU 12. We also note that in LBP-TOP 2D the textured image has been pose-normalized. The results can be worse for pure 2D-based recognition approaches. This proves that 3D face representation surpasses the 2D image representation by isolating the facial components such as expression, head pose and skin tone in this experiment. In our case, facial expression data can be extracted efficiently for our experiments. Contrariwise, to analyze expression in 2D, researchers need to carefully deal with challenges from pose prediction and skin color normalization, etc., and the error coming with the preprocessing stage can contaminate the training procedure. In short, we believe that the new data and our test approaches show promise. Further details are described in [41].

5.4. Case study — pain analysis with BP4D-spontaneous database

To further validate the usefulness of the 4D spontaneous data, we choose the spontaneous expression *physical pain* for a case study. The

Table 17
Spontaneous action unit recognition: confusion tables for best accuracies with nebula 4D approach.

Classified as →	AU1 offset	AU1 onset	No AU1
AU1 offset	23	5	35
AU1 onset	8	15	24
No AU1	18	7	98

Classified as →	AU2 offset	AU2 onset	No AU2
AU2 offset	14	2	35
AU2 onset	4	11	22
No AU2	11	1	113

Classified as →	AU4 offset	AU4 onset	No AU4
AU4 offset	17	0	39
AU4 onset	1	16	31
NO AU4	6	6	109

Classified as →	AU6 offset	AU6 onset	No AU6
AU6 offset	40	7	25
AU6 onset	9	36	17
No AU6	14	7	98

Classified as →	AU7 offset	AU7 onset	No AU7
AU7 offset	36	6	27
AU7 onset	12	18	25
No AU7	18	7	82

Classified as →	AU10 offset	AU10 onset	No AU10
AU10 offset	33	8	27
AU10 onset	3	27	25
NO AU10	10	5	94

Classified as →	AU12 offset	AU12 onset	No AU12
AU12 offset	27	6	34
AU12 onset	4	25	23
No AU12	22	6	78

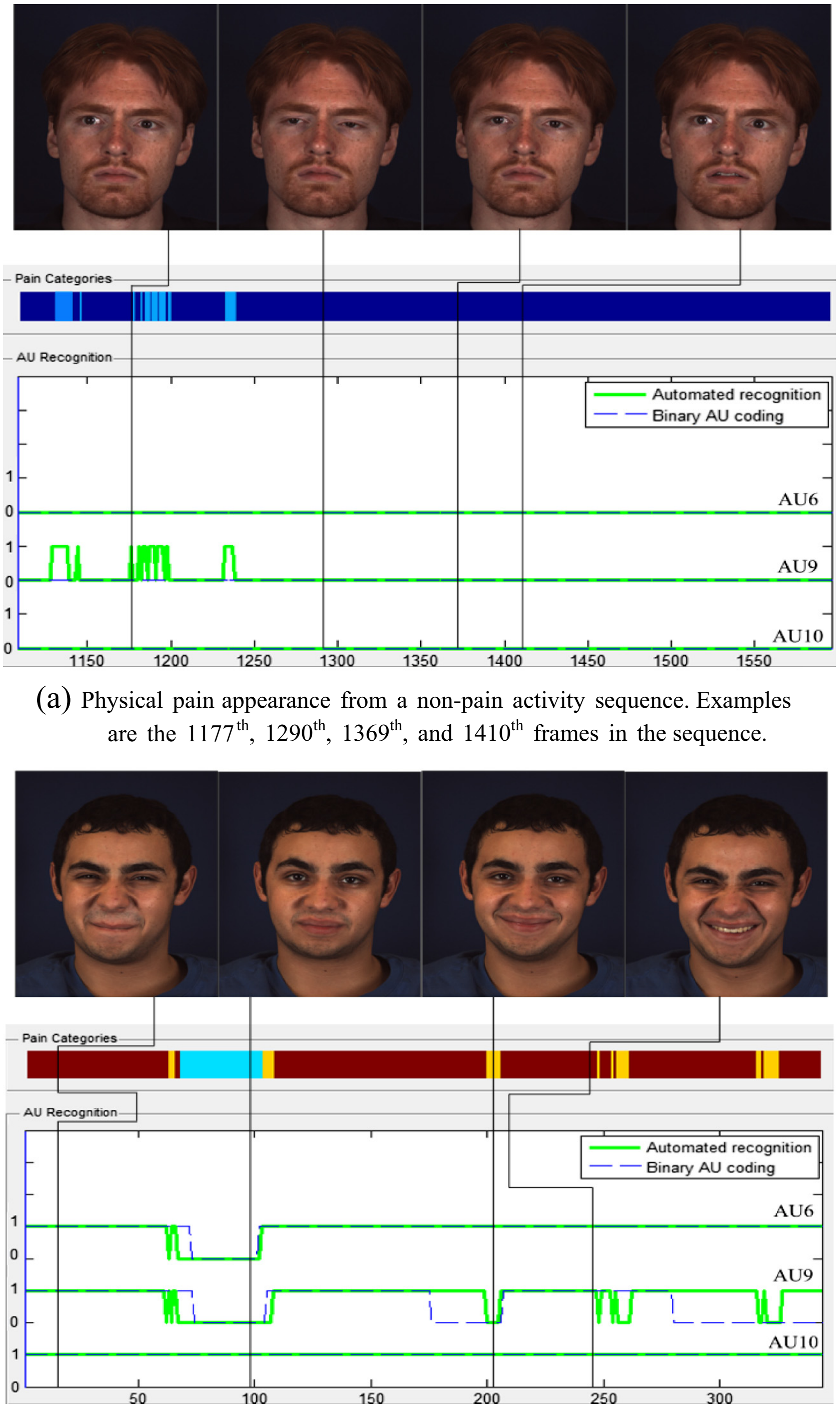
Classified as →	AU14 offset	AU14 onset	No AU14
AU14 offset	29	5	33
AU14 onset	5	23	25
No AU14	17	11	87

Classified as →	AU15 offset	AU15 onset	No AU15
AU15 offset	11	4	23
AU15 onset	5	10	22
NO AU15	4	5	119

Classified as →	AU17 offset	AU17 onset	No AU17
AU17 offset	46	5	37
AU17 onset	9	38	23
No AU17	18	6	103

Classified as →	AU23 offset	AU23 onset	No AU23
AU23 offset	19	6	31
AU23 onset	3	17	32
No AU23	12	7	109

Classified as →	AU24 offset	AU24 onset	No AU24
AU24 offset	10	0	36
AU24 onset	1	11	28
NO AU24	2	2	123



(a) Physical pain appearance from a non-pain activity sequence. Examples are the 1177th, 1290th, 1369th, and 1410th frames in the sequence.

(b) Physical pain appearance from a genuine pain sequence. Examples are the 14th, 99th, 203th, and 246th frames in the sequence.

Fig. 11. a: Physical pain appearance from a non-pain activity sequence. Examples are the 1177th, 1290th, 1369th, and 1410th frames in the sequence. b: Physical pain appearance from a genuine pain sequence. Examples are the 14th, 99th, 203th, and 246th frames in the sequence. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

goal is to evaluate spontaneous pain appearance with automated AU recognition. In this case, the contributed AUs may be different from the typical AUs for pain listed in Table 11. To find the contributed AUs, we use the paired *t*-test on the possible subset of pain AUs (AU 4, 6, 7, 9, 10) for all tasks of the database. This test is to assess whether pain and other spontaneous expressions are different in terms of the AU quantity. For each possible AU, the percentage of frames with this AU is reported for each task per subject. Here we find that in terms of AU quantity the differences between pain and all other expressions are significant with AU 6, AU 9, and AU 10 ($p < 0.05$, shown in Table 16). Therefore, the automated AU recognition system is built based on these three AUs. The positive data (AU appearance) set is acquired

Please cite this article as: X. Zhang, et al., BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database, Image Vis. Comput. (2014), http://dx.doi.org/10.1016/j.imavis.2014.06.002

from task 6 (a total of 19,524 frames), while the negative data (No AU appearance) is from task 2 (a total of 20,837 frames) as it has the minimum p value (4.50×10^{-7}).

To generate the facial features, face normalization and feature registration procedures are applied through rotating the face model to a frontal view, cropping the 3D face region, and normalizing it to a standard size of a 72×72 matrix along with the landmark and pose information. Then, a Gabor wavelet filter with 3 spatial scales and 8 orientations is applied to generate the feature vector.

Principal component analysis (PCA) is used to reduce the feature dimension to 200. Three binary support vector machines (SVMs) are then trained with one label per frame using a radial basis function (RBF) kernel for three AUs separately. The output margin to the SVM hyper plane is used to generate the ROC curves. A leave-one-subject-out validation method is used for all 41 subjects in the database. Each test subject's sequence from task 2 and task 6 has been used as the negative test sequence and positive test sequence, respectively.

The average AUC scores weighted by the positive sample numbers for AU 6, AU 9 and AU 10 are 0.865, 0.667, and 0.806 respectively. The pain appearance can be categorized in 8 ($=2^3$) types, depending on the occurrence of each AU (0 or 1). If none of the three AUs appears, the face shows no pain.

In Fig. 11(a–b), we illustrate the AU recognition performance, pain categories, and example frames. In each figure, sample frames are displayed on top. For AU 6, AU 9, and AU 10, each has been plotted with two curves, i.e., a green curve representing the automatic recognition result, and a dashed blue curve representing the ground truth from the binary AU coding. For each AU in a frame, it is labeled to 1 if the AU occurs, or 0 if the AU does not appear. Based on the eight kinds of combinations, different color indicates different pain appearances as shown in the bar "pain categories". 'Red" shows intense pain and "Blue" shows no pain.

Fig. 11a shows the result of pain detection along a sequence of task 2, in which the subject was watching a documentary. The accuracy for AU 6 and AU 10 is 1, and the AUC score for AU 9 is 0.93. In such a non-pain activity, the subject did not show any pain expression according to AU 6, AU 9 and AU 10. The results match the ground-truth (dashed blue lines) very well in most of the sequence except for very few frames where a single AU 9 is falsely identified.

Fig. 11b shows the pain detection results when the subject was submerging his hand in ice water (task 6). The AUC scores for AU 6 and AU 9 are 0.98 and 0.70, and the accuracy for AU 10 is 1. As an example, the second frame in Fig. 11b shows a less painful expression as compared to the others. Only AU 10 was detected for that frame.

Thus, we have taken physical pain as a study case to show how the spontaneous expression data and the corresponding meta-data (AU codes, pose, and feature data) could be used for automated expression recognition. Based on this study, more applications of 3D-based pain analysis and further investigation of a complete set of pain-related AUs could be developed in the future.

6. Conclusion and future work

In this paper, we reported our newly developed spontaneous 3D dynamic facial expression database (the so-called "BP4D-Spontaneous" — Binghamton–Pittsburgh 4D Spontaneous Facial Expression Database). Such a database can be a valuable resource to facilitate the research and development of human behavior analysis in security, HCI, psychology and biomedical applications.

It is worth noting that stimulating genuine emotions with spontaneous facial expressions in a controlled environment is still a challenging issue. Our BP4D-Spontaneous data focused on facial actions that are not deliberately posed. Rather, they occur in the course of social interaction and other social or non-social stimuli. The guided format using a professional actor and director as the experimenter sought to simulate a more natural setting while guaranteeing high quality 3D/2D dynamic facial expression data.

The AU annotation by expert coders provides valuable information about spontaneous facial behavior. However, the diversity of spontaneous expressions and the mixture of different expressions still pose big challenges for emotion and expression recognition.

Expression analysis on 2D data suffers from head pose variation, illumination change, self-occlusion, etc. These influences can be removed or reduced by 3D representations. BP4D-Spontaneous addresses this issue by providing 3D time-varying spontaneous facial expression data. When we use "Nebula" and LBP features to recognize AUs, we also noticed that the performance varies with different AUs. This may due to the different types of features exhibited by different AUs. For example, AU 4 and AU 9 create lines and furrows, while AU 25 reveals a new feature by showing teeth. That said, there may not be a panacea feature for recognition of all action units. In BP4D-Spontaneous, 3D/2D imaging data and the corresponding tracking points provide important information for feature distribution and discrimination. Combining the features from the 3D and 2D domains may improve the expression recognition performance.

In future work, other settings and image capturing setups might be considered. Data quality could be improved by using a wider range imaging system which is more robust to different illumination conditions. The database will also be expanded to include more subjects.

Moreover, our current database includes sequential geometric model data and texture data. In addition to the facial feature tracking algorithms, more powerful approaches need to be investigated in order to make the data processing and visualization fast and accurate. Automatic data annotation, registration, and efficient data representation (or compression) for micro-expression analysis will also be our next research direction.

Acknowledgments

This material is based upon the work supported in part by the National Science Foundation under grants IIS-1051103, IIS-1051169, CNS-1205664, and CNS-1205195. We would like to thank Nicki Siverling, Dean Rosenwald, and Shaun Zuratovic for FACS coding. We would also like to thank Dr. Peter Gerhardstein for the help on data collection.

References

[1] Man machine interaction group, http://www.mmifacedb.com/ 2005.
[2] T. Kanade, J.F. Cohn, Y. Tian, Comprehensive Database for Facial Expression Analysis, IEEE International Conference on Automatic Face and Gesture Recognition (FG), France, 2000.
[3] M. Pantic, L. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000).
[4] M. Pantic, L. Rothkrantz, Toward an affect-sensitive multimodal human-computer interaction, Proc. IEEE 91 (9) (2003).
[5] Y. Zhang, Q. Ji, Active and dynamic information fusion for facial expression understanding from image sequences, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (May 2005) 699–714.
[6] H. Wallbott, K. Scherer, Cues and channels in emotion recognition, J. Pers. Soc. Psychol. 51 (4) (1996) 690–699.
[7] E. Douglas-Cowie, R. Cowie, et al., The Humaine Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data, ACII, 2007, pp. 488–500.
[8] T. Anziger, K. Scherer, Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus, ACII, 2007.
[9] A. Savran, N. Alyuz, et al., Bosphorus Database for 3D Face Analysis, BIOID, 2008, pp. 47–56.
[10] J. Gross, R. Levenson, Emotion elicitation using films, Cogn. Emot. 9 (no. 1) (1995) 87–108.
[11] R. Cowie, R. Cornelius, Describing the emotional states that are expressed in speech, Speech Comm. 40 (1–2) (2003) 5–32.
[12] Di3D Inc., http://www.di3d.com.
[13] Z. Zeng, M. Pantic, G. Roisman, T. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.
[14] L. Yin, X. Wei, Y. Sun, W. J., M. Rosato, A 3D Facial Expression Database for Facial Behavior Research, IEEE Inter. Conf. on Automatic Face and Gesture Recognition, Southampton, UK, 2006, 2006.
[15] L. Yin, X. Chen, Y. Sun, T. Worm, M. Reale, A High-resolution 3D Dynamic Facial Expression Database, IEEE Inter. Conf. on Automatic Face and Gesture Recognition, Amsterdam, the Netherlands, Sept. 2008, Sept. 2008.

Please cite this article as: X. Zhang, et al., BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database, Image Vis. Comput. (2014), http://dx.doi.org/10.1016/j.imavis.2014.06.002

[16] P. Lucey, J.F. Cohn, J. Saragih, I. Matthews, Z. Ambadar, The Extended Cohn–Kanade Database: A Complete Facial Expression Database for Both Facial Action Units and Emotion Detection, IEEE CVPR4HB, 2010.

[17] J.F. Cohn, Z. Ambadar, P. Ekman, Observer-based Measurement of Facial Expression With the Facial Action Coding System, in: J.A. Coan, J.J.B. Allen (Eds.), The handbook of emotion elicitation and assessment, Oxford University Press Series in Affective Science Oxford University, New York, NY, 2007, pp. 203–221.

[18] J.F. Cohn, P. Ekman, Measuring Facial Action by Manual Coding, Facial EMG, and Automatic Facial Image Analysis, in: J.A. Harrigan, R. Rosenthal, K. Scherer (Eds.), Handbook of nonverbal behavior research methods in the affective sciences, Oxford University Press, NY, 2005, pp. 9–64.

[19] J. Jang, T. Kanade, Robust 3D Head Tracking by Online Feature Registration, Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, the Netherlands, 2008.

[20] I. Matthews, J. Xiao, S. Baker, 2D vs. 3D deformable face models: representational power, construction, and real-time fitting, Int. J. Comput. Vis. 75 (1) (2007) 93–113.

[21] L. Noldus, R. Trienes, A. Henriksen, H. Jansen, R. Jansen, The observer video-pro: new software for the collection, management, and presentation of time-structured data from videotapes and digital media files, Behav. Res. Methods Instrum. Comput. 32 (2000) 197–206.

[22] Z. Zeng, M. Pantic, G. Roisman, T. Huang, A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions, ACM International Conference on Multimodal Interfaces, 2007.

[23] RU-FACS, http://mplab.ucsd.edu/wordpress/?page_id=36 2013.

[24] G. Stratou, A. Ghosh, P. Debevec, L. Morency, Exploring the Effect of Illumination on Automatic Expression Recognition Using the ICT-3DRFE Database, IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2011.

[25] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, L. Van Gool, A 3D audio–visual corpus of affective communication, IEEE Trans. Multimedia 12 (6) (Oct. 2012) 591–598.

[26] S. Wang, Z. Liu, S. Lu, X. Wang, et al., A natural visible and infrared facial expression database for expression recognition and emotion inference, IEEE Trans. Multimedia 12 (7) (Nov. 2010) 682–691.

[27] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, J.F. Cohn, DISFA: a spontaneous facial action intensity database, IEEE Trans. Affect. Comput. 1 (2013).

[28] M. Mahoor, S. Cadavid, D. Messinger, J.F. Cohn, A Framework for Automated Measurement of the Intensity of Non-Posed Facial Action Units, 2nd IEEE Workshop on CVPR for Human communicative Behavior analysis (CVPR4HB), Miami Beach, June 25, 2009.

[29] J. Wang, L. Yin, X. Wei, Y. Sun, 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution, IEEE CVPR, 2006.

[30] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 681–685.

[31] M. Lyons, et al., Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (1999) 1357–1362.

[32] K. Schmidt, Z. Ambadar, J.F. Cohn, L. Reed, Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling, J. Nonverbal Behav. 30 (2006) 37–52.

[33] E. Goeleven, R. De Raedt, L. Leyman, B. Verschuere, The Karolinska directed emotional faces: a validation study, Cogn. Emot. 22 (6) (2008) 1094–1118.

[34] R. Gross, I. Matthews, J.F. Cohn, T. Kanade, S. Baker, Multi-PIE, Image Vis. Comput. 28 (5) (2010) 807–813.

[35] D. Cosker, E. Krumhuber, A. Hilton, A FACS Valid 3D Dynamic Action Unit Database With Applications to 3D Dynamic Morphable Facial Modeling, IEEE Inter. Conf. on Computer Vision (ICCV), 2011.

[36] P. Ekman, W. Friesen, J. Hager, Facial action coding system: Research Nexus, Network Research Information, Salt Lake City, 2002.

[37] J. Fleiss, Statistical Methods for Rates and Proportions, Wiley, NY, 1981.

[38] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable Model Fitting with a Mixture of Local Experts, IEEE Inter. Conf. on Computer Vision, 2009.

[39] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vis. 91 (2) (2011) 200–215.

[40] S. Canavan, X. Zhang, L. Yin, Fitting and Tracking 3D/4D Facial Data Using a Temporal Deformable Shape Model, IEEE International Conference on Multimedia &Expo (ICME13), San Jose, 2013, 2013.

[41] M. Reale, X. Zhang, L. Yin, Nebula Feature: A Space-time Feature for Posed and Spontaneous 4D Facial Behavior Analysis, IEEE Inter. Conf. on Automatic Face and Gesture Recognition (FG), 2013.

[42] Y. Sun, L. Yin, Facial Expression Recognition Based on 3D Dynamic Range Model Sequences, The 10th European Conference on Computer Vision (ECCV08), Oct. 2008, Marseille, France, Oct. 2008.

[43] J. Coan, J. Allen (Eds.), Oxford handbook on emotion elicitation and assessment, Oxford University Press, NY, 2007.

[44] C. Von Baeyer, T. Piira, C. Chambers, M. Trapannotto, L. Zeltzer, Guidelines for the cold pressor task as an experimental pain stimulus for use with children, J. Pain 6 (4) (2005) 218–227.

[45] G. Sandbach, S. Zafeiriou, M. Pantic, L. Yin, Static and dynamic 3D facial expression recognition: a comprehensive survey, Image Vis. Comput. 30 (10) (Oct. 2012) 683–697.

[46] D. Clark, On the induction of depressed mood in the laboratory: evaluation and comparison of the Velten and musical procedures, Adv. Behav. Res. Ther. 5 (1) (1983) 27–49.

[47] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (Jun. 2007) 915–928.

[48] Dacher Keltner, Evidence for the distinctness of embarrassment, shame, and guilt: a study of recalled antecedents and facial expressions of emotion, Cogn. Emot. 10.2 (1996) 155–172.

[49] Kenneth M. Prkachin, The consistency of facial expressions of pain: a comparison across modalities, Pain 51 (3) (1992) 297–306.

[50] Paul Ekman, Wallace V. Friesen, et al., Is the startle reaction an emotion? J. Pers. Soc. Psychol. (1985) 1416.

[51] Zara Ambadar, Jeffrey F. Cohn, Lawarence Ian Reed, All smiles are not created equal: morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous, J. Nonverbal Behav. 33 (1) (2009) 17–34.

[52] Charles Darwin, The Expression of the Emotions in Man and Animals, 3rd edition Oxford University, New York, 1872/1998.

[53] Richard J. Davidson, Paul Ekman, C.D. Saron, J.A. Senulis, Wallace V. Friesen, Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology I, J. Pers. Soc. Psychol. 58 (2) (1990) 330–341.

[54] Paul Ekman, Wallace V. Friesen, EMFACS: Coder's Instructions, Human Interaction Laboratory, University of California San Francisco, San Francisco, 1982.

[55] Paul Ekman, Erika Rosenberg, Joseph C. Hager, Facial Action Coding System Aect Interpretive Database (FACSAID), from http://nirc.com/Expression/FACSAID/facsaid.html 1998.

[56] Jeffrey M. Girard, Jeffrey F. Cohn, Mohammad H. Mahoor, Seyed M. Mavadati, Zakia Hammal, Dean Rosenwald, Social risk and depression: evidence from nonverbal behavior analysis, J. Image Vision Comput. (2014) (in press).

[57] Eddie Harmon-Jones, David M. Amodio, Leah R. Zinner, Social Psychological Methods of Emotion Elicitation, Handbook of emotion elicitation and assessment, Oxford University Press, NY, 2007.

[58] Mohammad E. Hoque, Daniel J. McDuff, Rosalind W. Picard, Exploring temporal patterns in classifying frustrated and delighted smiles, Affective Computing, IEEE Transactions on 3.3 (2012) 323–334.

[59] Patrick Lucey, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia Solomon, Iain Matthews, Painful data: The UNBC-McMaster shoulder pain expression archive database, J. Image Vision Comput. 30 (2012) 197–205.

[60] David K. Markus, The perception of “live” embarrassment: a social relations analysis of class presentations, Cogn. Emot. 13 (1) (1999) 105–117.

[61] Jonathan Rottenberg, Rebecca D. Ray, James J. Gross, Emotion Elicitation Using Film, in: J.A. Coan, J.J.B. Allen (Eds.), Handbook of emotion elicitation and assessment, Oxford University Press, NY, 2007, pp. 9–28.

[62] Paul Rozin, Jonathan Haidt, Clark R. McCauley, Disgust, in: M. Lewis, J.M. Haviland-Jones, L.F. Barrett (Eds.), Handbook of emotions, 3rd ed. Guilford Press, NY, 2008, pp. 757–776.

[63] J. Richard Landis, Gary G. Koch, The measurement of observer agreement for categorical data, Biometrics (1977) 159–174.

Please cite this article as: X. Zhang, et al., BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database, Image Vis. Comput. (2014), http://dx.doi.org/10.1016/j.imavis.2014.06.002