Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis

Zheng Zhang[1], Jeffrey M. Girard[2], Yue Wu[3], Xing Zhang[1], Peng Liu[1], Umur Ciftci[1], Shaun Canavan[1], Michael Reale[1], Andrew Horowitz[1], Huiyuan Yang[1], Jeffrey F. Cohn[2], Qiang Ji[3], and Lijun Yin*[1]

[1]Binghamton University, [2]University of Pittsburgh, [3]Rensselaer Polytechnic Institute

Abstract

Emotion is expressed in multiple modalities, yet most research has considered at most one or two. This stems in part from the lack of large, diverse, well-annotated, multimodal databases with which to develop and test algorithms. We present a well-annotated, multimodal, multidimensional spontaneous emotion corpus of 140 participants. Emotion inductions were highly varied. Data were acquired from a variety of sensors of the face that included high-resolution 3D dynamic imaging, high-resolution 2D video, and thermal (infrared) sensing, and contact physiological sensors that included electrical conductivity of the skin, respiration, blood pressure, and heart rate. Facial expression was annotated for both the occurrence and intensity of facial action units from 2D video by experts in the Facial Action Coding System (FACS). The corpus further includes derived features from 3D, 2D, and IR (infrared) sensors and baseline results for facial expression and action unit detection. The entire corpus will be made available to the research community.

1. Introduction

In the last 10 years, research on facial expression analysis has shifted its focus from posed behavior to non-posed (*i.e.*, spontaneous) behavior [34, 2, 14, 17, 23]. This shift has increased the difficulty of such analyses, but also their ecological validity and practical utility. In the next 10 years, a similar shift will occur from single modality to multimodal analyses, with increasing research integrating 2D and 3D videos, temperature dynamics, and physiological responses.

Researchers are already beginning to use 3D sensors and models to improve facial feature tracking and expression recognition [29, 34, 21, 25, 20, 33, 32]. However, because of the difficulty in collecting and labeling spontaneous behavior, these studies mainly focused on posed expressions.

Infrared imaging technology has also been employed for facial expression analysis [30, 15] due to its sensitivity to skin temperature and relative insensitivity to lighting conditions and skin color. However, existing work mainly utilized the temperature information as a single modality. Because the temperature distribution may not align well with facial appearance, it is challenging to extract temperature-based features for expression recognition.

Research has also shown the correlation of the physiological state to the emotion state of individuals [18, 24]. A number of databases have been developed successfully in recent years [13, 23, 1]. The utility of the physiological signals needs to be further investigated.

Complex human behavior can only be fully-understood by integrating physical features from multiple modalities (*e.g.*, facial expressions and physiological responses). Many studies have theoretically and empirically demonstrated the advantage of integrating multiple modalities in human emotion perception relative to using a single modality [18, 24, 13, 23]. However, the emotion-related modalities are typically studied separately.

To our knowledge, there is no database of emotional behavior that combines following multiple emotion related modalities: 2D and 3D face visual dynamics, skin temperature dynamics, and physiological responses.

Although there are several facial expression databases that include 3D data (*e.g.*, BU-3DFE [33], BU-4DFE [32], Bosphorus [21], ITC-3DRFE [25], ETH-3DAV [9], and 3D AU-DB [5]), they are all based on posed behavior and typically include few subjects, little diversity, limited ground truth labels, and limited metadata.

Recently, a 3D spontaneous facial expression database (BP4D) [35] with extensive labeling, metadata, and diversity was released to the research community. The 2D videos of this dataset were included in the second Facial Expression Recognition and Analysis Challenge (FERA) [28]. However, this dataset only includes 41 subjects, which lim-

*Main Contact: lijun@cs.binghamton.edu; Others: jeffcohn@pitt.edu; jiq@rpi.edu; realemj@sunyit.edu

its its statistic power and discriminative capacity for emotion classification.

Another recent database [1] includes multimodal data on body motion and electromyographic signals. However, this data is limited to multiple views other than the range data for study of chronic pain related emotions.

In short, as of yet, there is no corpus of sufficiently large size and ethnic diversity that includes the following information: 2D and 3D video of spontaneous facial behavior, thermal imaging, physiological data, expert FACS labels [8], and derivatives (*e.g.*, features).

These findings motivated us to develop a multimodal 3D dynamic spontaneous emotion corpus with metadata (*i.e.*, labels and feature derivatives). In this paper, we present a corpus that includes 140 subjects from various ethnic/racial ancestries: Black, White, Asian (including East-Asian and Middle-East-Asian), Hispanic/Latino, and others (*e.g.*, Native American). The emotion-related modalities include facial expressions, thermal, 2D and 3D dynamics, and physiological data.

Each subject experienced 10 tasks corresponding to 10 different emotion categories. The physiological data was collected by a vital sign sensor (*e.g.*, heart rate, blood pressure, respiration rate, skin conductivity (EDA)). The skin temperature was also collected by a thermal camera. To elicit authentic and ecologically-valid emotional expressions, we designed a protocol with four approaches integrated seamlessly, including social interview, film watching, physical experience, and controlled activities. A 3D dynamic imaging system is used to capture high-resolution 3D dynamic facial geometric data and video texture data. Such high-definition 3D dynamic (aka 4D) facial representation allows us to examine the fine structural change as well as the precise time course for spontaneous expressions. We have also processed and analyzed the dataset to provide a set of labels and feature derivatives in 2D/3D/IR in order to facilitate the utility of the new corpus. FACS codes (partial AUs) are annotated manually with respect to both their occurrence and intensity. The self-report and data validation have also been reported in the database.

The contribution of this work is three-fold:

1. This is the first multimodal data corpus with a large set of well-synchronized and aligned sensor modalities including high-definition 3D geometric facial sequences, 2D facial videos, thermal videos, physiological data sequences (heart rate, blood pressure, skin conductance (EDA), respiration rate).
2. The data is significantly expanded in terms of number of subjects with diverse ethnic/racial ancestries as compared to the existing databases. A procedure with 10 seamlessly-integrated tasks was applied by a professional performer/interviewer, resulting in the effective elicitation of spontaneous emotions.

Ethnic/Racial	Number	Proportion
Latino/Hispanic	14	10.0%
White	64	45.7%
African American	15	10.7%
Asian	46	32.9%
Others	1	0.7%

Table 1: Ethnic distribution across 140 participants.

3. A large set of metadata was created, including feature points from 2D videos, 3D videos, and thermal videos, head pose, etc. FACS AUs were encoded in terms of their occurrence and intensity. With 140 subjects included in the database, there are over 10TB high quality data generated for the research community. The data have been verified and validated through a number of applications, including expression classification, AU detection, and thermal data classification.

The remainder of this paper gives details about the data acquisition, organization, annotation, and validation.

2. Data Acquisition

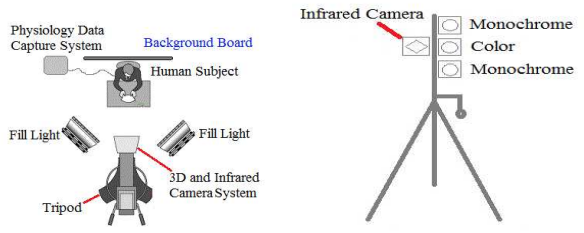
2.1. Participants

140 subjects have been recruited to participate in data collection at the Binghamton University. There are 58 males and 82 females, with ages ranging from 18 to 66 years old. Ethnic/Racial Ancestries include Black, White, Asian (including East-Asian and Middle-East-Asian), Hispanic/Latino, and others (*e.g.*, Native American). Table 1 shows the ethnic distribution. Following the IRB approved protocol, the informed consent form was signed by each subject before the start of data collection.

2.2. Recording System Setup and Synchronization

The data capture system (as shown in Figure 1a) includes a 3D dynamic imaging system, a thermal signal sensor, and a physiological signal sensor system. The 3D dynamic imaging system (Di3D[1]) contains a 3D stereo imaging sensor and a 2D video sensor. The thermal sensor is mounted on the 3D dynamic imaging system with a tripod in a fixed position (as shown in Figure 1b), and all these sensors are positioned in the same distance to a subject. The physiological signals are collected using the Biopac[2] MP150 system. It captures vital sign signals in a very high sample rates, including blood pressure, respiration rate, heart rate and electrodermal activity (EDA). Detailed configurations are depicted in the following subsections. Note that the system synchronization is critical for data collection from various modality sensors. Due to each sensor has its own

[1]http://www.di4d.com
[2]http://www.biopac.com



(a) System aerial view (b) Relative position of cameras

Figure 1: Recording system

machine to control, we developed a program to trigger the recording from the start to the end across all three sensors simultaneously. This is realized through the control of a master machine by sending a trigger signal to three sensors concurrently.

2.2.1 3D dynamic imaging system

The 3D face model sequences and their corresponding 2D texture image sequences are captured by the Di3D dynamic imaging system. Associated with two symmetric lights, the system is composed of a 3D sensor with a pair of stereo monochrome cameras aligned vertically and an RGB 2D color camera placed in between the stereo cameras. The 3D model of each frame is created by the dense passive stereo photogrammetry method. Each facial model contains about 30k – 50k vertices giving much detailed geometric information at RMS accuracy of 0.2 mm. And the resolution of each 2D texture image is 1040×1392 pixels. Consider the trade off between emotion granularity and computing complexity, we set the frame rate to 25fps. This is also consistent to the video rate of the thermal sensor.

2.2.2 Thermal sensor

The thermal camera that we used is FLIR[3] A655sc Longwave infrared camera. This camera captured thermal videos in resolution of 640×480 per frame with $25°$ Lens and 17 micron pixels with temperature range of -40 and $150°$C. The frame rate is 50 fps with the full resolution of 640×480. The spectral range is $7.5 - 14.0\mu$m. In order to better synchronize all sensors in our system, we set the capture rate of the thermal sensor to 25 fps.

2.2.3 Physiological signal sensing system

The physiological data were collected by Biopac MP150 data acquisition system. Its measurement capacity is in the range [-25mmHg, 300mmHg] for blood pressure, [0, 200 breaths/minutes] for respiration rate, and [30, 300 beats/minute] for heart rate. The blood pressure signal (mmHg) is captured through noninvasive blood pressure (Biopac NIBP100D) monitoring system containing two units, finger unit and an inflatable cuff. The finger unit captures data from an index finger and a middle finger of a hand and an inflatable cuff is placed on an arm for calibration. Having the blood pressure signal with peak count, the pulse rate (beat/minute), systolic blood pressure (mmHg), and diastolic blood pressure (mmHg) are derived. The respiration signal (measured in voltage) is captured by a respiration belt wearing around the chest. Given the peak count, it derives the parameter - respiration rate (breaths/minute).

The electrodermal activity (EDA) (measured in micro Siemens) is captured through two leads placed on a right palm connecting a wrist watch. The EDA signal is the indication of arousal level with various skin conductivity.

In general, the system captures physiological signals in a very high sample rate at 1000Hz. The resulting data include heart rate, respiration rate, systolic blood pressure, diastolic blood pressure, and electrodermal activity (EDA).

2.3. Emotion Elicitation

In order to evoke a range of authentic emotions in a laboratory environment, we designed a protocol of ten tasks with seamless transitions. Motivated by the work [35], a professional actor was hired to host the entire interview procedure during data collection. Interviews with a skilled interviewer can elicit a wide range of emotional expressions and interpersonal behavior.

Four methods were employed in the protocol, which include interpersonal conversation, film clip watching, cold pressor, and designed physical experiences. Ten activities (tasks T1–T10 as shown in Table 2) were conducted with a natural transition from positive emotions to negative emotions. Between any two tasks, there was a brief pause for self-report.

Data collection started with a social interview in which the interviewer told a joke for a relaxed and amused atmosphere. Then the subject's 3D avatar was created on-site and displayed to the subject for a surprising effect. A negative feeling was then elicited by showing the subject a video clip of a 911 emergency call, followed by a sudden burst of sound for a startled expression. After a pause, the interviewer posed a question to induce a skeptical expression, followed by an embarrassment induction by asking the subject to do a silly performance. Then a fearful feeling was generated through a dart game experience, followed by a physical discomfort experience by having the subject submerge a hand into ice water. After that, the interviewer induced an upset feeling in the subject by pretending to complain about the subject's poor performance on the ice water task. Finally, an unpleasant odor was presented to the subject to induce a disgusted feeling.

Note that this emotion elicitation protocol has more tasks

[3] http://www.flir.com/

Task	Activity	Target Emotion
T1	Interview: Listen to a funny joke	Happiness Amusement
T2	Graphic show: Watch 3D avatar of participant	Surprise
T3	Video clip: 911 emergency phone call	Sadness
T4	Experience a sudden burst of sound	Startle Surprise
T5	Interview: True or false question	Skeptical
T6	Improvise a silly song	Embarrassment
T7	Experience physical threat in dart game	Fear Nervous
T8	Cold pressor: Submerge hand into ice water	Physical pain
T9	Interview: Complained for a poor performance	Angry
T10	Experience smelly odor	Disgust

Table 2: Ten tasks for spontaneous emotion elicitation.

than the other reported methods. According to the compound emotions theory [7], the surprised feeling could be positive or negative. In our experiment, a fearful surprise was triggered by a siren in T4 and a joyful surprise was induced by seeing a self 3D face in T2. We treat them in two different categories.

2.4. Self Report

As stated in Section 2.3, immediately after each task, every participant was provided with a short period to report the feeling that he/she had experienced.

A tablet was used to choose emotions and their intensities from a list of possible choices (*e.g.*, relaxed, surprised, sad, happy/amusement, skeptical, physical pain, disgusted, embarrassed, nervous, scared/fear, angry/upset, frustrated, and startled/shocked). 5-point Likert-type scales from "very slightly" to "extremely" were used to rate the emotion intensity. Participants were allowed to choose multiple emotion categories as well as to input other emotion categories if none of provided options fit their experience.

Among the data collected, we conducted statistical tests on all ten tasks. The top three emotions voted by all participants of each task are displayed in Figure 2. As seen in Figure 2, the majority vote of each task fits well with the emotion that the task was intended to elicit, which demonstrates that the designed elicitation protocol was effective.

3. Database Organization

The new corpus is structured by participants. Each participant is associated with 10 tasks including high-resolution 3D model sequences, 2D RGB videos, thermal

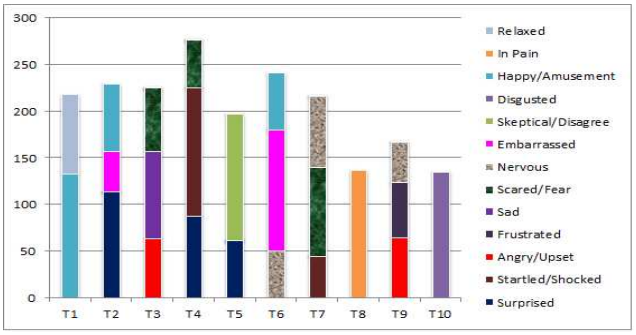


Figure 2: Emotion distribution from self-report.

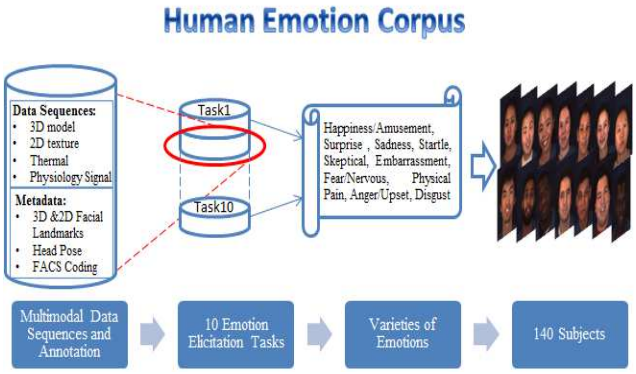


Figure 3: Database overall structure

videos, and sequences of physiological signals (*i.e.*, respiration rate, blood pressure, EDA, and heart rate). Figure 3 shows the overall structure of database. The average size of each subject is about 100GB, resulting in over 10TB with about 1.4 million frames in total. Figure 4 illustrates sample data sequences of four modalities from a subject.

In addition, the metadata are also generated, including manually labeled action units (both occurrence and intensity) on four tasks, automatically tracked head poses, and 3D/2D/IR facial landmarks. Detailed annotations and method will be described in the next section.

4. Data Annotation and Descriptive Statistics

4.1. FACS Coding

Expert FACS coders annotated facial action units during four tasks (*i.e.*, happiness/amusement, embarrassment, fear/nervous, and physical pain) for all 140 subjects. Therefore, we have 560 (*i.e.*, 4×140) data sessions coded.

4.1.1 AU occurrence

Segments of the most facially-expressive 20 seconds of each task and a total of 34 facial action units were occurrence coded by five expert FACS coders. Coders annotated onsets when the action units reached the B-level of intensity (as defined by the FACS manual) and offsets when they

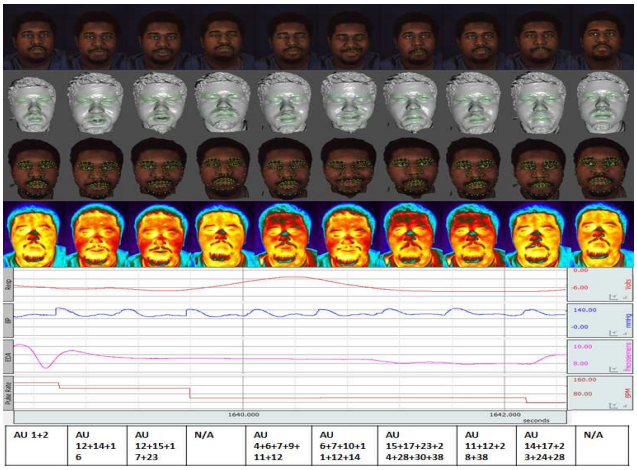
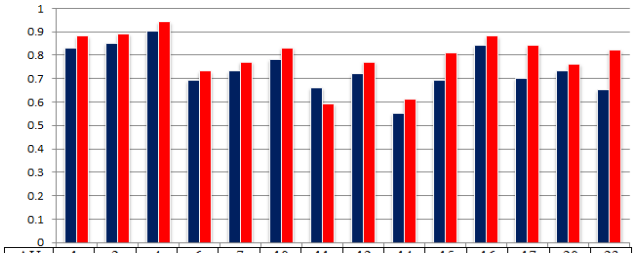


Figure 4: Sample data sequences from a participant including original 2D texture (first row), shaded model (second row), textured model (third row), thermal image (fourth row), physiology signal(fifth row: respiration rate, blood pressure, EDA, heart rate) and corresponding action units(last row).



AU	1	2	4	6	7	10	11	12	14	15	16	17	20	23
MCC	0.83	0.85	0.9	0.69	0.73	0.78	0.66	0.72	0.55	0.69	0.84	0.70	0.73	0.65
S	0.88	0.89	0.94	0.73	0.77	0.83	0.59	0.77	0.61	0.81	0.88	0.84	0.76	0.82

Figure 5: Occurrence reliability with two kinds of metrics.

dropped below it. Table 3 shows the base rate for all 34 action units.

To assess inter-rater reliability, 94 sessions were randomly selected for comparison coding. Two or more of the five coders coded these videos. Across the action units with base rates higher than 5%, the mean value (S) as seen from Figure 5 was 0.79, ranging from 0.59 for AU 11 to 0.94 for AU 4. Among all chance-adjusted reliability indices, the S index is robust to the most problems (*e.g.*, skewed base rates) and is especially suited to binary occurrence coding [37]. According to Altman's benchmarks, these results indicate very good reliability (>0.8) for 8 action units, good reliability (>.6)for 5 action units, and moderate reliability (>.4)for 1 action unit. For consistency with the past, results are also presented in Figure 5 using the overall Matthew's Correlation Coefficient (MCC).

AU	BR	Events	Frames	AU	BR	Events	Frames
1	10%	514	19083	20	15%	808	29197
2	8%	438	16145	22	2%	425	4160
4	6%	374	11419	23	17%	1614	32941
5	1%	59	1352	24	4%	331	7699
6	50%	774	98021	27	0%	6	179
7	66%	804	130693	28	2%	141	4631
9	4%	223	7046	29	0%	0	0
10	65%	848	127636	30	3%	189	5084
11	41%	833	80936	31	0%	22	544
12	58%	773	113704	32	0%	41	933
13	0%	14	329	33	0%	6	141
14	60%	1044	118533	34	0%	2	48
15	11%	865	21132	35	0%	0	0
16	33%	1828	64794	36	0%	0	0
17	13%	1216	25576	37	0%	7	66
18	0%	66	982	38	1%	163	2487
19	1%	104	1332	39	1%	77	1005

Table 3: AU occurrence for 140 subjects(BR = base rate).

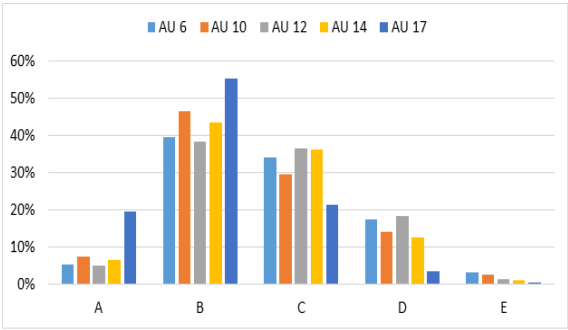


Figure 6: Percentage of frames at each intensity level.

4.1.2 AU intensity

AUs 6, 10, 12, 14, and 17 were intensity coded for a subset from the whole database. Coding was completed by two expert coders. The distribution of intensity levels was similar across action units, with B-level frames being the most common, followed by C-level frames, D-level frames, A-level frames, and E-level frames, in descending order. Although occurrence coders delimited events at the B-level, intensity coders annotated additional frames before and after each event. Many of these additional frames were A-level frames.

Percentage of frames at each intensity level is illustrated in Figure 6. Across the action units that were coded for intensity, the mean inter-rater reliability (weighted S) was 0.76, ranging from 0.70 for AU 6 to 0.84 for AUs 10 and 12 (Table 4). Although many interval-level performance metrics (*e.g.*, PCC, ICC, and MSE) have been used to calculate the reliability of intensity coding, intensity codes are ordinal in nature and require a categorical reliability metric. Here, we apply ordinal weighting to the S index. According to Altman's benchmarks, these results indicate good reliability for three AUs and very good reliability for two AUs.

AU	06	10	12	14	17
Weighted S	0.7	0.84	0.84	0.71	0.71

Table 4: Intensity reliability across 5 action units.

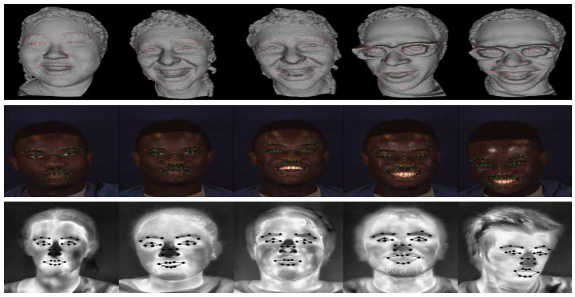


Figure 7: Samples of feature points tracking on 3D models (1st row: 83 feature points tracked by SI-SSM), 2D texture sequences (2nd row: 49 feature points tracked by *zface*), and thermal frames (3rd row: 28 feature points).

This analysis is based on redundant coding of 9 sessions by both coders; the sessions selected differ between AUs.

4.2. Feature Point Tracking

4.2.1 3D feature tracking

To track feature points from the 3D geometric face surface directly, we apply a so-called shape index-based statistical shape model (SI-SSM) [3] for such a task. Similar to [3], 83 landmark points are defined on the face surface including the eyes, nose, mouth, eyebrows, and face contour. Utilizing the SI-SSM allows us to track a range of dynamic expressions and model transformations (translation, rotation, etc.), by making use of both the global and local shapes of the input face model. The global face shape is constructed by a parameterized model S_G from a set of training data, each with 83 patches centered at 83 feature landmarks. Similarly, the local face shape S_L is represented by the shape index values of each patch. PCA is applied to both the global and local shape models to learn the modes of variation from the training data. Both the global and local feature vectors are then combined into one model $S_{GL} = \{S_G, S_L\}$, representing the face surface shape with expression deformation more adaptively.

To detect and track features on 3D face models, the classic cross correlation template matching scheme is applied to compute the correlation score between the each patch of the SI-SSM and the input mesh model patches. We compared the tracked features from the SI-SSM approach to manually annotated ground truth resulting in a mean squared error of 2.5. Example tracked frames can be seen in Figure 7 (first row).

4.2.2 2D feature tracking

Two-dimensional facial expression sequences were automatically tracked using the *zface* software [12]. By applying cascade regression to person-independent 3D registration (inferred from the 2D video), *zface* tracked 49 facial feature points with various head poses in each video frame. Using this approach, facial feature points remain invariant across head pose over a range of approximately 60 degrees. Figure 7 (second row) shows several sample frames with tracked points.

4.2.3 Thermal feature tracking

In order to make three modalities face data (3D, 2D, IR (Infra-red)) easy to align each other, we have also tracked the 28 facial landmarks from the thermal temperature data directly.

Initially we pre-process the thermal temperature data to increase their local contrast. Then the Constrained Local Model (CLM) [6] is applied to sequentially perform independent facial landmark detection and global refinement based on the face shape pattern constraint. In detection, we first initialize the facial landmark locations using the mean face shape based on the initialized thermal eye locations. Then, we search for each facial landmark independently in the local region with the Gabor wavelet, phase-based displacement estimation method [38], and a pre-built offline feature databases. Given the independent facial landmark detection results, the face shape is refined with the Active Shape Model [4]. In tracking, the facial landmarks are initialized as the locations in the last framework. For local independent landmark searching, both the online template from the last frame and the offline databases are combined for better prediction. ASM is also applied to refine the independent detection results. We tested the thermal facial landmark tracking approach on the thermal sequences of the database, and calculated the landmark tracking error as the distance between the tracked landmark locations and the manually annotated ground truth landmark locations, which is normalized by the inter-ocular distance. If we consider the images with error less than 10% of the inter-ocular distance as successfully detected images, the detection rate is 91.57%. Figure 7 (third row) shows the tracking result on a sample set of thermal images with different participants.

4.3. Head Pose

Head pose is an important cue for understanding emotional expressions. It is tracked and included in the database as one of the meta-data. Three orientations (yaw, roll and pitch) are estimated through video sequences. Since the head pose information can be derived directly from our tracked 3D points on the 3D face model sequences, here we

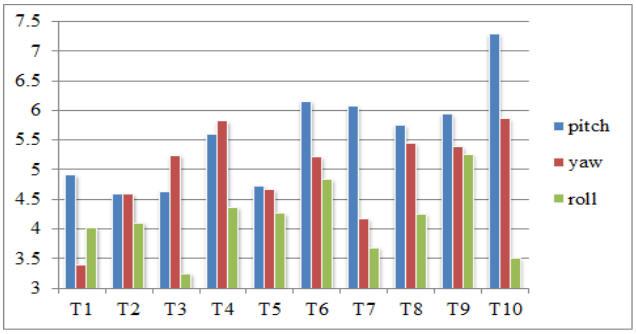


Figure 8: Head pose variations under different tasks.

focus on tracking the head pose from 2D videos. A cylindrical head tracker [11] is used to get head pose from 2D videos. The tracker works person-independently, and has concurrent validity with 2D+3D AAM [16] with magnetic motion capture device [11].

We randomly select 60 subjects for the statistical analysis. As shown in Table 6a, over 90% of frames are less than 10 degree with respect to the front view. To show the head pose variations with different emotions, we compute the standard derivations of all 60 subjects across 10 tasks. In Figure 8, vertical axis stands for standard derivation of head pose. Among all 10 tasks, tasks T6-T10 have clearly larger pose variations in pitch than the other tasks. Except for T3-T4, the pitch variation is more dramatic than roll and yaw. Except for T2, the roll variation appears to be the smallest among the three orientations, meaning that the rolling head is not common nor comfortable for exhibiting emotions. Having such a finding, pitch could be used for disclosing more clues on emotional status than the other head orientations. The physical experience method in the process of emotion elicitation could have more dramatic head motions than the other emotion induction methods.

5. Experiments and Validation

5.1. 3D Dynamic Spontaneous Expression Analysis

To validate the usefulness of the data, we applied the approach reported by Huang *et al*. [10] to adapt a generic model to the model sequences. The hybrid approach using two vertex mapping algorithms, displacement mapping and point-to-surface mapping, and a regional blending algorithm are used to reconstruct the facial surface detail. The adapted models have the same number of vertices across the corresponding 3D video sequence. Thereby, the vertex correspondence across the range model sequence is established. The vertices on the adapted model can be tracked by finding the displacement of tracked vertices of two neighboring frames.

In order to make it comparable to the state of the art [35, 26], we implemented the 3D dynamic facial expression descriptor based on primitive feature labels and HMM classifier as described in [26]. 60 subjects were selected for performance evaluation. Among ten tasks, six tasks (T1, T3, T4, T7, T9, and T10, corresponding to six prototype expressions) of each subject were used for classification. A 10-fold cross validation procedure was applied with 90% subjects for training and 10% subjects for testing. The result shows that the average correct recognition rate is 74.8%. As comparison, the same approach was applied to the BP4D database [35], resulting in a 73.7% recognition rate for classifying spontaneous 3D dynamic expressions (joy, anger, surprise, disgust, fear, and sadness) on average. We have also applied the same approach to 3D dynamic posed expressions using a public database BU-4DFE [32], where 81.2% recognition rate was achieved for classifying six posed expressions. Apparently, our new dataset has the comparable quality to the BP4D database in terms of the 3D modality. The 3D dynamic spontaneous facial expressions show much more variety and subtlety in appearance and timing than the posed expressions. This still poses a big challenge for analysis of naturally occurred facial behavior.

5.2. Facial Expression Analysis on Thermal Data

To validate the utility of the thermal data, we have also conducted experiments on facial expression recognition using the thermal videos of randamly selected 60 subjects. We applied the thermal video descriptor reported in [15] for such a task. The face region of each thermal video clip is warped to the frontal view based on scale-invariant feature transform(SIFT) flow, generating a corresponding SIFT flow video clip. Then, the thermal video cuboids are segmented from each thermal video clip based on max pooling and motion video cuboids are segmented from each SIFT flow video clip based on average pooling. Thermal video words and motion video words are clustered by k-means cluster. Finally, each video is represented by a histogram of the bag of SIFT Flow and facial temperature changes video words. The resulting histogram is used as a descriptor for classification by the support vector machine(SVM). The recognition accuracy is 91%.

For comparison, we have also applied a state-of-the-art approach reported in [31] to test on our database. The features derived from the temperature difference matrix on forehead, left cheek and right cheek were used. The recognition accuracy is 62%. We further compared the two methods on randomly selected 22 subjects from USTC-NVIE database [30] and achieved recognition accuracy 71% and 59%, respectively, on classifying six prototype expressions.

5.3. Task Classification on Physiological Data

To validate the utility of the collected physiological data, we conducted emotion recognition experiments based on those data. In our experiments, we randomly selected 45

subjects, and classify the 10 tasks (emotions) from the set of features [18] extracted from the physiological signals, including 8 features from EDA (*i.e.*, mean and variance of the normalized signal (Mn, Vn), mean and RMS of the 1st derivative (Md, RMSd), average rising time and recover time of SCRs, average of negative derivative, and their proportion to all derivative values), and 7 features from blood pressure signals (*i.e.*, Mn, Vn, Md, RMSd, and pulse rate, diastolic blood pressure, systolic blood pressure variance), and 5 features from respiration signals (*i.e.*, Mn, Vn, Md, RMSd, and respiration rate variance).

In our first experiment we have selected five tasks (Tasks 1, 3, 4, 7, and 10) which target happiness, sadness, startle, fear and disgust emotions. Using 10-fold cross validation and RBF kernel SVM, the average accuracy of five-class emotion recognition is 59.5%. Moreover, we mapped the emotion classes into binary classes of low and high arousal using the emotion semantic space described in [22]. Based on the new classes, we used the same classifier and achieved 60.5% accuracy in classifying 10 tasks (emotions) from our database.

5.4. AU Detection and Recognition

We performed experiments to detect and recognize FACS Action Units based on 3D dynamic facial model sequences. We developed a log-normal (LN) based 4D polynomial fitting approach for generating spatio-temporal features [19]. Given the 3D dynamic model sequences, depth information is extracted for each frame, and the neighborhood around each spatio-temporal point is fit to a 4D polynomial (using the best-fitting log-normal function to model temporal behavior). The dynamic curvature values, the static curvature values, and the shape index values are computed from each feature, and histogram for each spatial region is formed from the corresponding features. More details of the algorithm (the LN-based 4D feature approach) can be found in [19].

Action units are tested for individually in subsequences (31 frames long) extracted from the full video sequences. Positive samples are subsequences that contain a single AU preceded and followed by the absence of that particular AU. For classification, the Leave-One-Subject-Out approach was employed using SVM. Subsequences for 7 AUs were extracted, which resulted in 213 sample subsequences extracted from a subset of the database. The method automatically finds the "best" 7-frame window in a 31-frame subsequence (referred to LN(31)). We also list the results from using only the 7-frame windows containing the AUs on both our approach (referred to as LN(7)) and LBP-TOP [36] (referred to as LBP(7)). The results are presented in Table 5. We also performed a cross-database test by training on data extracted from the BP4D-Spontaneous database [35] and testing on the new data. AUs 1, 2, 6, 16, and 17 were tested. The results can be seen in Table 6b. These experiments again used the 31-frame subsequences with the LN(31) approach.

	AUC			F1		
AU	*LBP(7)*	*LN(7)*	*LN(31)*	*LBP(7)*	*LN(7)*	*LN(31)*
1	.926	.868	.860	.817	.817	.725
2	1.000	1.000	1.000	.750	.873	.733
6	1.000	1.000	.920	.899	1.000	.899
8	.944	1.000	.972	.916	.829	.916
14	.838	.949	.865	.802	.918	.849
16	.952	.824	.927	.853	.790	.852
17	.891	.828	.701	.786	.786	.666
WA	.895	.904	.857	.818	.859	.803

Table 5: AU Depth Data Results (WA = Weighted Average based on AU sequence counts)

Angle	Pitch	Yaw	Roll
$< 5°$	72.8	64.5	82.6
$< 10°$	92.7	92.2	97.6
$< 15°$	98.7	98.1	99.6
$< 20°$	99.5	99.6	99.9

(a) Proportion of frames with certain range of rotation angles

AU	Acc.	AUC	F1
1	.818	.876	.817
2	.875	.938	.873
6	.900	.920	.899
16	.765	.875	.765
17	.714	.780	.714
WA	.776	.850	.776

(b) AU Depth Cross-Database Results

Table 6: Statistics of head pose and AU recognition

6. Conclusion and Future Work

In this paper, we have presented a new multimodal spontaneous emotion database (MMSE) for the research community in order to facilitate the research of the field. We have employed the state-of-the-art algorithms to label and validate the data. Partial data have also been used successfully for application of video-based heart rate estimation [27]. However, our current work has certain limitations, which give rise to our future work as follows: (1) we will expand the database from several aspects, including more subjects, AUs, and intensity coding; (2) we will also extract the physiological features and study the cross-correlation of multimodal data and their fusion schemes. As a result, the database will be sustained and updated progressively by including all the derivatives for the research community.

Acknowledgement

This material is based upon the work supported in part by the National Science Foundation under grants CNS-1205664 and CNS-1205195. We would like to thank Nicki Siverling for technical assistance. We would also like to thank Dr. Peter Gerhardstein for his help in data collection.

References

[1] M. S. Aung, S. Kaltwang, B. Romera-Paredes, M. Pantic, et al. The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset. *IEEE Trans. on Affective Computing*, 2015. 1, 2

[2] M. S. Bartlett, G. C. Littlewort, et al. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006. 1

[3] S. Canavan, L. Yin, et al. Landmark localization on 3d/4d range data using a shape index-based statistical shape model with global and local constraints. *Computer Vision and Image Understanding (CVIU)*, 139:136–148, 2015. 6

[4] T. F. Cootes, C. J. Taylor, et al. Active shape models—their training and application. *CVIU*, 1995. 6

[5] D. Cosker, E. Krumhuber, and A. Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *ICCV*, 2011. 1

[6] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054 – 3067, 2008. 6

[7] S. Du, Y. Tao, and A. Martinez. Compound facial expressions of emotion. *Proc. of the NAS*, 111(15), 2014. 4

[8] P. Ekman and W. V. Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978. 2

[9] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Trans. on Multimedia*, 2010. 1

[10] Y. Huang, X. Zhang, L. Yin, et al. Reshaping 3d facial scans for facial appearance modeling and 3d facial expression analysis. *Image and Vision Computing*, 30(10), 2012. 7

[11] J. S. Jang and T. Kanade. Robust 3d head tracking by online feature registration. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2008. 7

[12] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015. 6

[13] S. Koelstra, C. Mühl, M. Soleymani, et al. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. on Affective Computing*, 3(1):18–31, 2012. 1

[14] G. C. Littlewort, M. S. Bartlett, and K. Lee. Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, 27(12), 2009. 1

[15] P. Liu and L. Yin. Spontaneous facial expression analysis based on temperature changes and head motions. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2015. 1, 7

[16] I. Matthews, J. Xiao, and S. Baker. 2d vs. 3d deformable face models: Representational power, construction, and real-time fitting. *International Journal of Computer Vision*, 2007. 7

[17] G. McKeown, M. Valstar, et al. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. on Affective Computing*, 3(1):5–17, 2012. 1

[18] R. W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. on PAMI*, 23(10):1175–1191, 2001. 1, 8

[19] M. Reale. Automatic analysis of facial activity for multimodal human-machine applications. *PhD dissertation, Binghamton University*, 2014. 8

[20] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10), 2012. 1

[21] A. Savran, N. Alyüz, H. Dibeklioğlu, et al. Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, pages 47–56. 2008. 1

[22] K. R. Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005. 8

[23] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. on Affective Computing*, 3(1):42–55, 2012. 1

[24] G. Stemmler. Methodological considerations in the psychophysiological study of emotion. *Handbook of affective sciences*, pages 225–255, 2003. 1

[25] G. Stratou, A. Ghosh, et al. Exploring the effect of illumination on automatic expression recognition using the ict-3drfe database. *Image and Vision Computing*, 30(10), 2012. 1

[26] Y. Sun and L. Yin. Facial expression recognition based on 3d dynamic range model sequences. In *ECCV*. 2008. 7

[27] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe. Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *CVPR*, 2016. 8

[28] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn. FERA'15 2nd facial expression recognition and analysis challenge. In *FG*, 2015. 1

[29] J. Wang, L. Yin, et al. 3d facial expression recognition based on primitive surface feature distribution. In *CVPR*, 2006. 1

[30] S. Wang, Z. Liu, et al. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Trans. on Multimedia*, 12(7), 2010. 1, 7

[31] S. Wang, Z. Liu, Z. Wang, et al. Analyses of a multimodal spontaneous facial expression database. *IEEE Trans. on Affective Computing*, 4(1):34–46, 2013. 7

[32] L. Yin, X. Chen, et al. A high-resolution 3d dynamic facial expression database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2008. 1, 7

[33] L. Yin, X. Wei, et al. A 3d facial expression database for facial behavior research. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2006. 1

[34] Z. Zeng, M. Pantic, G. Roisman, T. S. Huang, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. on PAMI*, 31(1), 2009. 1

[35] X. Zhang, L. Yin, J. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. BP4D-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32, 2014. 1, 3, 7, 8

[36] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. on PAMI*, 29(6):915–928, 2007. 8

[37] X. Zhao et al. 19 assumptions behind intercoder reliability indices. *Communication Yearbook 36*, 36:419, 2012. 5

[38] Z. Zhu and Q. Ji. Robust pose invariant facial feature detection and tracking in real-time. In *ICPR*, 2006. 6