Received 31 January 2023, accepted 8 February 2023, date of publication 16 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3246047

RESEARCH ARTICLE

KTFEv2: Multimodal Facial Emotion Database and Its Analysis

HUNG NGUYEN[1], NHA TRAN[1], HIEN D. NGUYEN[2,3], LOAN NGUYEN[4], AND KAZUNORI KOTANI[5]

[1]Faculty of Information Technology, Ho Chi Minh University of Education, Ho Chi Minh City 700000, Vietnam
[2]Faculty of Computer Science, University of Information Technology, Ho Chi Minh City 729110, Vietnam
[3]Vietnam National University Ho Chi Minh City, Ho Chi Minh City 700890, Vietnam
[4]School of Accounting, University of Economics Ho Chi Minh City, Ho Chi Minh City 743000, Vietnam
[5]School of Information, Japan Advanced Institute of Science and Technology, Nomi 923-1211, Japan

Corresponding author: Nha Tran (nhatt@hcmue.edu.vn)

This work was supported in part by the Ministry of Education and Training of Vietnam under Grant B2019-SPS-01.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Science's Faculty Council.

ABSTRACT In recent years, the focus of human emotion analysis has gradually shifted towards not only using visible information, but also thermal infrared (IR) information. This requires a great deal of facial emotion data both in visible and thermal IR information. However, most existing databases contain either visible information or posed thermal IR information only. For these reasons, we propose and establish a multimodal facial emotion database including both natural spontaneous visible and thermal IR videos. Beside updating more thermal infrared information, the built dataset in this study also enhances the information of intensity emotions. In which, each emotion is classified into three levels (low, medium, and high). Seven spontaneous emotions from thirty subjects are recorded in the database. Audio and visual stimuli were used to elicit emotions during the experiment. After the standard procedure of collecting data finished, the database has been through the careful annotation and verification procedure. Furthermore, the built database is analyzed by using modern machine learning models such as CNN, ResNet50, YOLO, and using a combination of different models to analyze the dataset. The obtained results are feasible and show that this dataset is useful for use in practice. The results of thermal data analysis provide us with a promising idea for future research on estimating human emotion.

INDEX TERMS Facial expression, facial emotion, thermal image, visible image, spontaneous database, KTFE database.

I. INTRODUCTION

Although human emotion plays a central part in our lives and there is much research focused on facial emotion analysis since the work of Darwin in 1872 [1], There is no clear explanation for how the human brain analyzes facial emotions and how computers can achieve the same accuracy rate of automatic facial emotion analysis as humans. Progress in the field of human emotion analysis is significant for the development of psychology as a scientific discipline. In another way, the research of facial emotion relies on knowledge about human emotional experience, as well as the relation between emotional experience and affective expression [2].

Besides, according to the temperature, everything emits infrared energy, called heat. The heat signature of an object is the infrared radiation it emits. In general, an object emits more radiation the hotter it is [3]. A thermal image, commonly referred to as a thermal camera, is essentially a heat sensor that can pick up on even the smallest temperature variations [4]. The apparatus gathers infrared radiation from nearby objects and builds an electronic representation of the scene using data on temperature differences [5]. For example,

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

for classifying facial emotions, skin temperature variation is useful [6] and facial expressions are good behavior related to emotions [7].

Nowadays, there are a few data containing both expression and emotion which were revealed from real experiments. Moreover, to study human emotional experience and expression in more detail and on a scientific level, and to develop and benchmark methods for automatic facial emotion analysis, researchers are in need of rich sets of data. Therefore, to overcome the related gaps and to contribute to the need of emotional databases, we have recorded a multimodal database which contains the participant responses via thermal infrared (IR) camera. The database is freely available to the academic community.

In recent years, There are many researches focused on facial expression analysis [36], [37], [38], [39], [40], [41], [42], [43], [44], such as Mohan et al. [36] proposed deep convolutional neural networks (DCNN) have two branches, the score-level fusion technique is adopted to compute the final classification score on five benchmark available databases consisting of seven basic emotions. The obtained results demonstrate that the proposed method outperforms all state-of-the-art methods on all the databases. The study in [37] proposed FER-net, which is a convolution neural network to distinguish efficiently. In this architecture, features are fed into a softmax classifier for identifying facial expressions. This method achieved high performance comparable to some of the state-of-the-art methods. Due to the tremendous attention from researchers of facial expression analysis innumerable natural databases have been built.

However, those databases do not meet criteria for application in the practice. USTC- NVIE database, which was collected by Key Laboratory of Computing and Communication Software of Anhui Province (CCSL), labeled many emotions of more than 200 objects [8]. Each emotion only has one level, thus it does not describe practical human emotion clearly. KTFE database was collected from thermal Facial Emotion Recognition (FER) pairs [9]. Nonetheless, the thermal file in that dataset is a specification file which only can be read by a copyrighted software coming with the camera. Thus, it is very difficult for users to use that dataset.

In this study, the details of materials, methods to design and collect data, data annotation, and data verification are given. This paper proposes a standard method for collecting the multimodal facial emotion database. This database, called KTFEv2, includes both natural spontaneous visible and thermal videos, the increasing number of collected objects. Additionally, the KTFEv2 database improves the information about emotional intensity; in which, each emotion of an object is classified into three levels: low, medium, and high. These improvements help the collected database to satisfy requirements to apply in the real-world. Besides, The built dataset has been analyzed by modern models of machine learning for emotion recognition, such as CNN, ResNet50, YOLO, and combined models and obtained effective results for using it. KTFEv2 database is simple to use, it contains the temperature matrix CSV file, the visible, and thermal images have been annotated. The results on thermal data show that this dataset allows research on facial expressions and emotions to have more realistic approaches.

II. REVIEW OF EXISTENT NATURAL AND INFRARED DATABASES

Emotion analysis and human-machine interaction are increasingly attracting interest, the number of databases created to serve the research on emotion recognition is increasing such as The Japanese Female Facial Expression (JAFFE) [10], CK [11], CK+ [12], FER2013 [13], DFEW [14]. However, most common visual databases are often created in the laboratory, so emotions are most likely to be posed and over-expressed such as JAFFE. The JAFFE (Japanese Female Facial Expression) database is a collection of 213 images of 7 facial expressions (happiness, surprise, sadness, anger, disgust, fear, and neutral) performed by 10 Japanese female models.

With FER2013, and DFEW database is built to collect facial expression images or videos from the Internet. Data collection often conflicts with privacy or ethics, and it both time consuming and error in the labeling process. The images in the FER2013 database vary in resolution, lighting, and other factors. The DFEW database contains a variety of difficult interferences, including intense lighting, occlusions, and erratic pose changes, which can affect the performance of facial expression recognition algorithms. The information and brief comparison of those data are shown in several comprehensive surveys [15], [16], [17], [18], [19].

The temperature distribution over faces and the vein branches are captured using infrared thermal images, are not sensitive to imaging conditions. Thus, thermal facial emotion estimation is widely used in biomedical applications. There are very few thermal facial databases developed to support research on human facial expressions and emotions. Only a very small number of thermal face databases are available in the literature as compared to the number of visible databases that already exist. Additionally, those databases only contain a small number of spontaneous thermal data and a few posed thermal data. In this document, the current popular databases of infrared facial expressions are listed and compared. The information used in comparison includes the name, the number of subjects, the wave band of thermal camera, elicitation, and expression description as Table 1.

In addition, there are several other thermal databases such as [22], [23], and [24]. However, the availability of thermal image face databases is limited: there is a lower demand for thermal images than for 2d databases, and thermal images are significantly more difficult to obtain [15].

The USTC-NVIE database is one of the most commonly used. However, their data collection procedure for inducing emotions contains an error. The gaps between each emotion clip in their video clips to elicit emotion are 1-2 minutes long, which is too short for participants to establish a neutral emotion status. They make no mention of the recording

TABLE 1. Current popular thermal facial database.

Name	Size	Wave band	Elicitation	Expression description
NIST Equinox [20]	600 subjects	8-12 μm 3-5 μm	Posed	Smiling, frowning, surprise.
IRIS [21]	30 subjects	7-14 μm	Posed	Surprise, laughing, anger.
USTC-NVIE [8]	215 subjects	8-14 μm	Posed, Spontaneous	Anger, disgust, fear, surprise, happiness, sadness.
KTFE [9]	26 subject	8-14 μm	Spontaneous	Anger, disgust, fear, surprise, happiness, sadness, neutral

time before the end of each emotion clip. Human tempers change later than emotions. As a result, the time remaining before the end of each emotion clip is very important [9]. The KTFE database is also one of the most commonly used databases. However, the number of participants is 26 and thermal information files are SVX files not widely supported. These reasons motivated us to propose and build up another natural visible and infrared facial emotion database.

TABLE 2. Participant information.

Number	Age	Sex	Education	Glasses	Nationality
2	32	2 Male	PostDoc	2 No	Vietnamese
1	31	1 Male	PostDoc	1 No	Vietnamese
2	30	1 Male, 1 Female	Phd	1 Yes, 1 No	Vietnamese
1	29	1 Male	Master	1 Yes	Vietnamese
6	28	3 Male, 3 Female	3Master, 3Phd	5 Yes, 1 No	Vietnamese, Thai
1	27	1 Male	Master	1 No	Vietnamese
5	26	3 Male, 2 Female	4Master, IPhd	3 Yes, 2 No	Vietnamese
5	25	4 Male, 1 Female	2Master, 3Phd	2 Yes, 3 No	Vietnamese, Chinese
6	24	4 Female, 2 Male	Bachelor	4 Yes, 2No	Vietnamese, Thai
1	12	1 Male	Pupil	1 No	Japanese
Total: 30					

III. MATERIALS AND METHOD

A. PARTICIPANT RECRUITMENT AND PREPARATION

Thirty participants ranging in age from 11 to 32 years old were recruited. They come from Vietnam, Japan, China, and Thailand. The majority of those chosen were JAIST students (Japan Advanced Institute of Science and Technology). See Table 2 for the breakdown of participants gender, nationality, educational level, glasses wearing status and age.

Participants were given the consent forms upon arriving at the data collection site and were asked to provide written consent after fully reading the form indicating that they are willing to participate in data collection.

After being given a date and time to participate in data collection, all participants were instructed to rest, keep a positive attitude for two hours prior to the measurements, and refrain from applying any cosmetics to their faces during the experiment. Before taking the data, they were explained carefully about the data collection procedure and its purpose.

B. MEASUREMENT DEVICES AND ENVIRONMENT

1) ROOM SETUP

The experiment room is located on the eighth floor of an isolated area. It has L shaped with 8m × 12m × 3.5m and the omitted area is about 6m2. The room was always kept silent to prevent any effect from evoking the participant's emotions. Due to the facial surface's sensitivity to the ambient temperature, the room's temperature is kept between 24 and 26 degrees Celsius throughout the experiment. The air conditioning system of the building is used to regulate the room's temperature and humidity. The flow of air conditioning was not directed to the testing area. Both the door and the curtains were closed throughout the experiment to maintain the constant lighting between day and night. In addition to an infrared camera, the experiment space was furnished with a laptop, desk, chair, LCD screen, mass storage disk, headphones, and two unique curtains. Two curtains divided the participant from the experimenter, making the participants feel more at ease and less shy, which made it simpler to elicit their emotions [6].

2) CAMERA SETUP

An Infrared Camera NEC R300 is utilized to capture the visible and thermal videos. The infrared camera has a long wavelength infrared (LWIR) camera that operates from 8 to 14 m in the infrared spectrum and a visible camera with 3.1 megapixels and 5 frames per second. The thermal sensitivity is 0.030C at 300C. Thermal infrared imaging data were captured at 5ft/s. In front of the participants, the camera was positioned 0.85 meters away and 1.5 meters above the floor. Before each experiment, a calibration was established, and it was updated automatically once every minute to get the participants' accurate temperatures. The NS9500 PRO and NS9500 STD are used to view, enhance, analyze, and extract the thermal data respectively and NS9500 PRO is used to capture both visible and thermal data.. It also supports a real-time monitor.

C. PROCEDURES

1) STIMULI

In this experiment, we use carefully chosen emotional video clips to arouse the participant's emotions. Four people obtained the online video clips, which the authors then judged. There are four angry clips, four disgusting clips, four fearful clips, four fearful games, six happy clips, seven sad clips, three surprised clips, and two neutral clips. Additionally, each emotion class is divided into four subclasses based on the intensive levels.

2) DATA ACQUISITION

During data collection, there was only one experimenter and one participant in the laboratory. the Participant was seated comfortably in a chair in front of a laptop screen. The data acquisition procedure fixes the former database mistake. Depending on the participants, we did not ask them to keep their glasses on or take them off. In order to capture spontaneous emotions, we did not also demand that they maintain a fixed angle of gaze. Prior to collecting data, the participants received an explanation of the experiment's goals and steps, and they were then instructed to put on headphones. Instrumental music was played before and after each session to assist the participant in regaining a neutral state of mind. We only tested one emotion per session, paying particular attention to the time lag phenomenon. We ended this session to uphold human rights when participants refused to record some terrified or furious clips. Each person underwent testing, and after that, we requested self-reports from them and their input for each emotion video clip. The recorded videos were given labels thanks to these self-reported data.

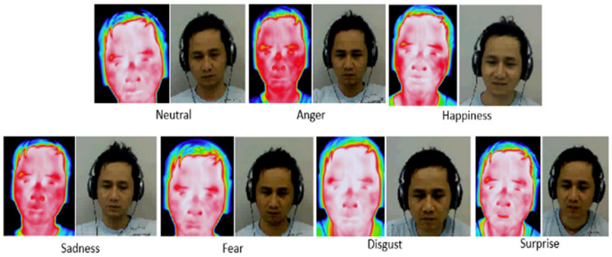


FIGURE 1. Examples of seven emotions' thermal and visible images.

D. KTFEv2 DATABASE DESIGN

The second version of KTFE database, called KTFEv2 database, which contains seven spontaneous emotions of 30 subjects, includes 214 gigabytes of visible and thermal facial emotion videos, visible facial expression image database, thermal facial expression image database, fusion of visible and thermal database, and thermal sequence database. To obtain the thermal expression image database, we manually choose the expressions using NS9500STD software. There are three persons to select manually the suitable frames for every emotion of each person and extract into thermal images. The visible image database is also manually extracted and chosen by two persons. Fig.1 shows the sample thermal and visible images of seven emotions.

IV. ANALYSIS OF EFFECTIVENESS OF ELICITING

In this section, we evaluate a method of emotion elicitation. Using self-report of participants about their emotions, we estimate the effectiveness of emotion-elicitation video clips.

A. METHODOLOGY

In the database acquisition process, after each process, participants will make a self-report about their feelings. In each self-report, they will answer some questions which are designed by us. For example, ''Have you ever watched these clips?'', ''What are your feelings (picking up from 7 emotions)?'', ''In 5-levels of feeling, which scale are you feeling? and so on. From the most viewed clips, we choose four clips for sadness, four clips for fear, four clips for disgust, four clips for happiness, four clips for sadness and three clips for surprise. We calculate the mean evaluation values of each emotion as well as the valence and arousal, which reflect the overall evaluation results for each emotion eliciting video clips [25].

B. RESULTS AND ANALYSIS

From Fig. 2 to Fig. 7, the mean of participant self-report data for each emotion are shown. We have some conclusion from those results: Firstly, according to those pictures, most of the video clips induce the desired emotions. Therefore, our clips almost work well and effectively. Another cue to support the effectiveness of our clips is the means of the valences. With happiness, surprise, the positive emotions have positive meanings of the valences, and the negative emotions, sadness, anger, disgust, fear have negative meanings of the valences. Secondly, the positive means of arousal prove that all video clips, induced emotion, are almost successful.

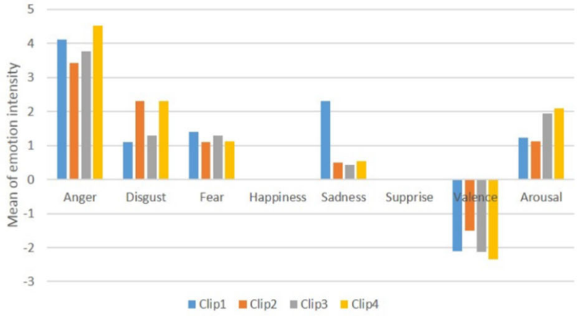


FIGURE 2. Video used to induce anger.

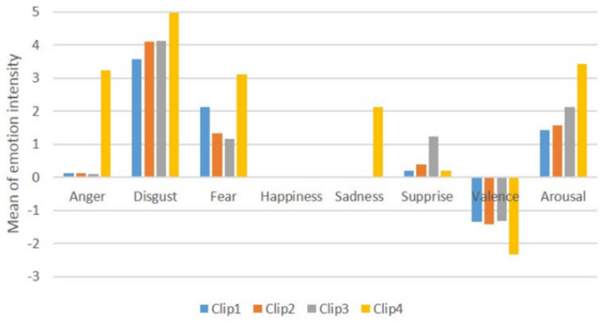


FIGURE 3. Video used to induce disgust.

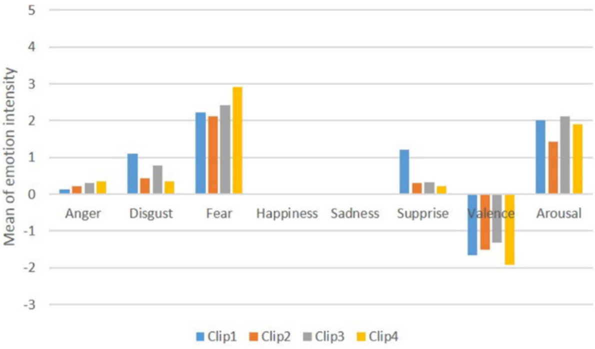


FIGURE 4. Video used to induce fear.

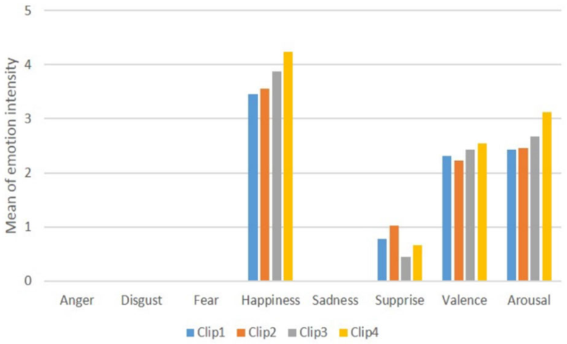


FIGURE 5. Video used to induce happiness.

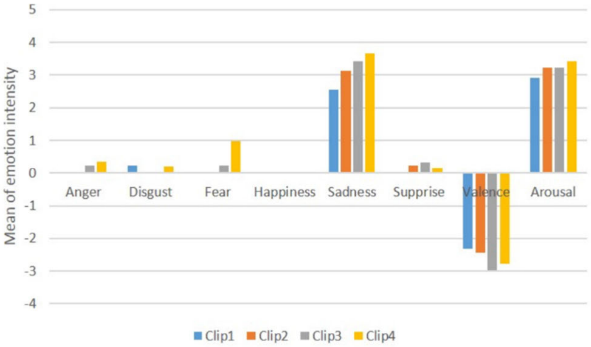


FIGURE 6. Video used to induce sadness.

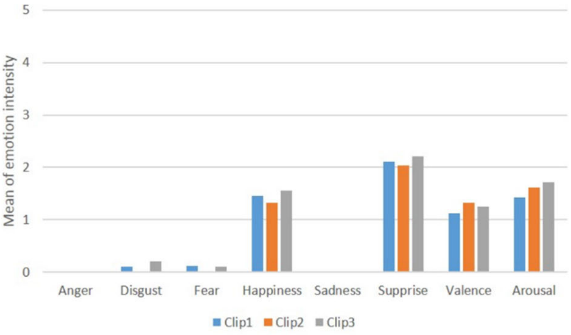


FIGURE 7. Video used to induce surprise.

From Fig.2 to Fig.7, we can infer that the video clips used to evoke an emotion could induce multiple emotions. For example, the clips for anger can evoke some degree of disgust, fear, and sad emotions. The clips for disgust can indicate some degree of anger, fear, and surprise. The clips for happiness can evoke some degree of surprise. This is consistent with the previous study results described in [25] and [26]. From those phenomena, we can develop new complex emotions which are not limited to six basic categories. They also give some prior information to support estimating human emotion.

V. ESTIMATION OF HUMAN EMOTION

A. DATA ANNOTATION

1) ANNOTATION BY SPONTANEOUS EMOTION (SE)

The database contained seven spontaneous emotions of 30 subjects. The number of objects associated with each emotion varies. There are 160 videos selected to label each emotion and obtain a visible image, thermal image, and temperature matrix CSV files by using an application the infRec Analyzer NS9500 Professional (NS9500 PRO), it is a highly functional software that can perform the measurement, analysis, and real-time output report of thermal and visible images. The data extraction process is shown in Fig. 8.

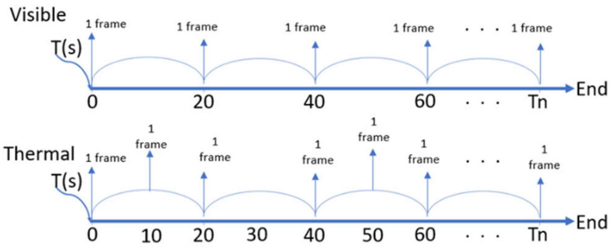


FIGURE 8. Diagram describing image extraction from videos.

The visible image starts at time t=0 and takes 1 frame, then every 20 seconds (corresponding to 100 frames) we take 1 frame.

The thermal image starts at time t=0 and takes 1 frame, then every 10 seconds (corresponding to 50 frames) we take 1 frame.

After extracting image data from the original database video, we proceed to assign labels for the database evaluation experiment (Fig. 9).

For training the YOLO network, we use input images and an annotation file in .txt format. The YOLO annotation file will have a structure of <id-class><center-x><center-y> <bbox-width><bbox-height>. Therefore, for visible images, we use Viola-Jones algorithm to recognize faces and then save the bounding box values.

We did the above calculations to get the result of <center-x> <center-y><bbox-width><bbox-height> which matches the correct YOLO annotation file format. The result is shown in Fig. 10.

However, for thermal images, it is difficult to use Viola-Jones algorithm or face detection algorithms on visible

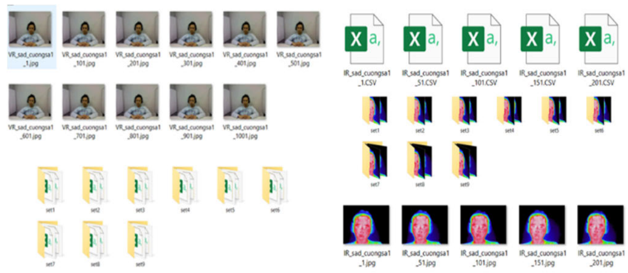
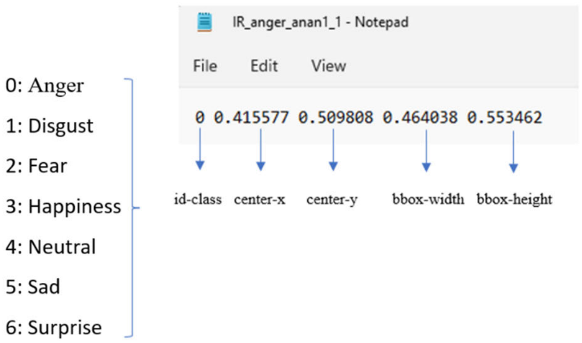


FIGURE 9. Archive data that have been extracted from videos.



0: Anger
1: Disgust
2: Fear
3: Happiness
4: Neutral
5: Sad
6: Surprise

IR_anger_anan1_1 - Notepad

File Edit View

```
0 0.415577 0.509808 0.464038 0.553462
```

FIGURE 10. File *.txt with the same name is generated for each image file.

images, so we have performed manual labeling by Roboflow, as shown Fig. 11.

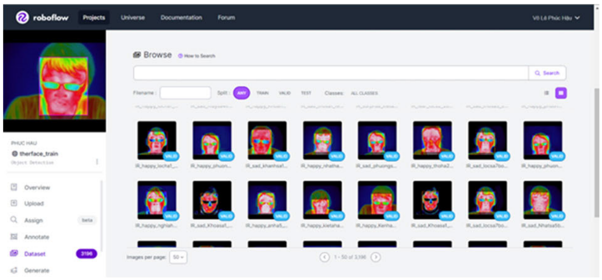


FIGURE 11. The image is labeled in Roboflow.

2) ANNOTATION BY SPONTANEOUS INTENSITY EMOTION (SIE)

In addition to the facial emotion dataset, with KTFEv2 emotions are expressed with various intensity. With a large-size database, it is difficult to experiment with feature extraction, classification, and ranking, so we divided the videos into smaller segments, extracting the highest intensity frames (apex) from the image data. visible and then get the corresponding frame from the thermal image data. Continuing the same process with varying degrees of intensity, labeling of emotion type and intensity was performed for each frame of all data. The process of performing the above labeling involves the participation of members of the research team and the use of software that supports NS9500 PRO. Fig. 12, shows an example of a labeled happiness emotional intensity.

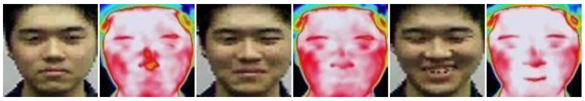
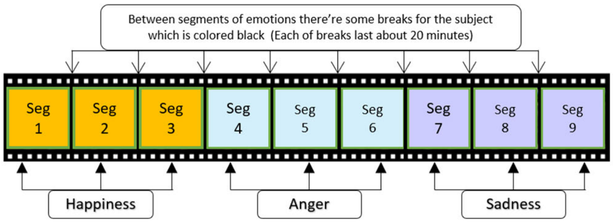


FIGURE 12. Example of of spontaneous intensity emotion - SIE (From left to right, the level of happiness is low, medium, and high) [33].

An average of 48 minutes and 32 seconds is the runtime of each video. We divided the video into sections (Fig. 13). The length of each section is 2-4 minutes. There are breaks for the subject in between segments of emotions; each break lasts about 20 minutes and is highlighted in black (in this period, the difference between emotional intensities is very low).



Between segments of emotions there're some breaks for the subject which is colored black (Each of breaks last about 20 minutes)
Seg 1
Seg 2
Seg 3
Seg 4
Seg 5
Seg 6
Seg 7
Seg 8
Seg 9
Happiness
Anger
Sadness

FIGURE 13. Design of the experimental procedure.

We store videos in folders after separating them into segments of varying intensities. First, we separate the segments according to the emotions of the subjects: happy, angry, and sad. Second, within each emotion folder, we continue to divide the segment into three categories: low, medium, and high. Finally, we separate the segments for thermal and fusion segments. Images with both thermal and visible frames are referred to as fusion segments. We categorize our segments into three folder layers, which are emotions, intensity, and image type.

B. VALIDATION

In order to ensure objectivity and high reliability for labeled data, the survey subjects were teachers, students, staff of universities, and educational centers aged from 19 to 25 years old. For one emotion, we selected 3 candidates in the database and took their pictures to question in the survey. Each candidate has 3 pictures in one emotion, therefore, we ask the volunteers to observe the images, and emotional validation, and rank the rate of that emotion based on their opinions. The sample results of this survey are displayed in Fig.14.

We developed a survey and invited various individuals to judge and offer opinions based on their experience and observation in order to evaluate the SIE database. We provide the survey participants with a list of inquiries based on the emotion-segmentation images. Based on their observations, the volunteers' opinions serve as the basis for the questions' answers. In order to create the survey questions, we select three candidates from the database for each emotion and take their pictures. We ask the volunteers to view the images and rank the rate of each candidate's respective emotion based on their assessments because each candidate has three images

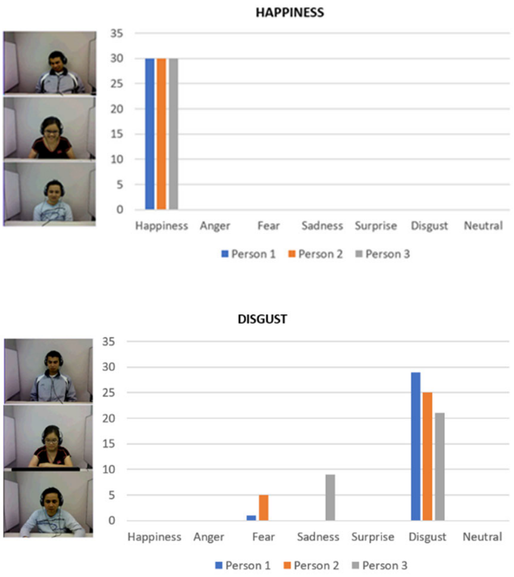


FIGURE 14. Survey answers for SE.

representing that emotion. Fig.15 displays a sample of the survey's result.

The findings of the survey led to some of the following conclusions:

- SE database: The results show the percentage of survey participants who totally agree with the assigned label for happiness and neutral emotions. Besides these 2 feelings, the surveyor said that the remaining emotion labeled is not really true from their point of view. They predicted another emotion during the survey. Therefore, using visible images to identify emotions is not enough information, it is necessary to have other supporting data sources such as thermal images.
- SIE database: angry emotions are difficult to discern the intensity through image, it's more confusing between levels than sadness and happiness. Happiness has the highest percentage matching the level label. For sadness, the results are slightly less accurate than for happiness.

C. ESTIMATION USING SPONTANEOUS EMOTION DATABASE (SE)

To evaluate the feasibility of the labeled dataset, we design four experiments on deep learning models such as CNN and ResNet50, and YOLO for data sets divided by emotion. Besides, the t-ROI, HOG feature extraction method, and SVM, SVM+ classification models are also used. In addition, the experiment combines two types of data: visible image and thermal image.

1) CONVOLUTIONAL NEURAL NETWORK AND ResNet50 MODEL

The dataset is extracted from a small part of the image database above, and the dataset of images is shown in Table 3. The training dataset will be divided into 2 parts, 80 training, and 20 testing. Besides, to increase the data source, we apply the data augmentation technique through basic transformations such as horizontal inversion (Flip) and image rotation in many angles (Rotation).

In this experiment (Exp-No.1), we propose a simple Convolutional Neural network (CNN), and pre-train model ResNet50 to train the model. CNN is a deep learning model that gives good results in classification problems with

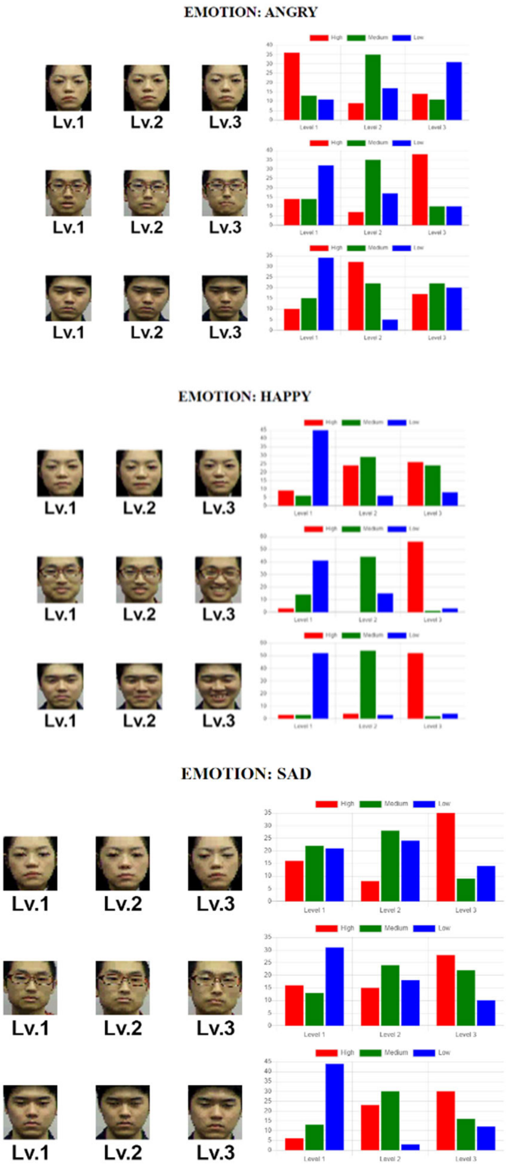


FIGURE 15. Survey answers for SIE [68].

TABLE 3. The number of images of each emotion in the dataset.

DATA	AN	DI	FE	HA	NE	SA	SU	TOTAL
Training set	744	360	386	552	360	1408	416	4208
Test set	186	90	92	138	90	352	104	1052
Total	930	450	460	690	450	1760	520	5260

automatic feature extraction. With the CNN model, we design a model consisting of 5 layers CONV-POOL, as shown in Fig. 16 with an input image normalized to 120 × 120.

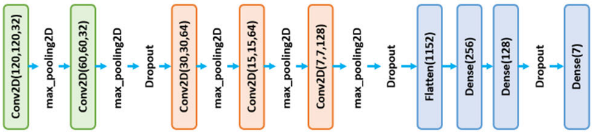


FIGURE 16. Emotion estimation by the CNN model.

ResNet50 [27] is a deep convolutional neural network with 50 layers that aims to solve the vanishing gradient problem. Connect a shortcut that "skips" some layers, converting the regular network to a residual network. It trained on over a million images from the ImageNet database. Therefore, we also use the ResNet50 network to experiment with the dataset using transfer learning. The data is normalized to 224 × 224 size to fit the input of the ResNet50 network.

TABLE 4. Experimental results with CNN.

Model	Metric	An	Dis	Fe	Ha	Ne	Sa	Su
Vi_ CNN	Precision	76	62	83	87	90	89	83
	Recall	85	97	51	89	71	85	68
	F1-score	80	76	63	88	80	87	75
Ther CNN	Precision	90	93	86	93	97	89	88
	Recall	97	83	91	87	91	97	83
	F1-score	93	87	88	90	94	93	85

TABLE 5. Experimental results with ResNet50 model.

Model	Metric	An	Dis	Fe	Ha	Ne	Sa	Su
Vi_ ResNet	Precision	76	62	83	87	90	89	83
	Recall	85	97	51	89	71	85	68
	F1-score	80	76	63	88	80	87	75
Th_ ResNet	Precision	90	93	86	93	97	89	88
	Recall	97	83	91	87	91	97	83
	F1-score	93	87	88	90	94	93	85

Table 4, Table 5, Table 6 shows that the CNN and ResNet50 models on the thermal image dataset have higher accuracy than the visible image. RestNet50 that model gives better results on both visible and thermal images.

From the experimental results Exp.01, we continue to experiment (Exp02) with the manual feature extraction method and apply the combined models technique. The t-ROI method [28] extracts features for thermal images. Then we use the SVM model for classification. For visible images, the HOG method is used to extract features and classify them by the SVM model. The SVM+ method [29] is used to combine the two models. The temperature information is privileged information, used only during model training.

In Table 7, Table 8, the accuracy of the results was on average 77.2 for visible images, 94.9 for thermal images, and 86.8 for combined models.

TABLE 6. Compare the results of the CNN model with ResNet50.

Model	CNN	ResNet50
Visible	79.60	82.87
Thermal	90.58	93.23

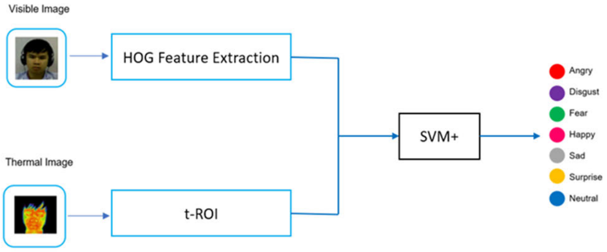


FIGURE 17. Model fusion using t-ROI, HOG feature with SVM+.

TABLE 7. Experimental results with HOG, t-ROI, SVM, and SVM+.

Model	Metric	An	Dis	Fe	Ha	Ne	Sa	Su
Ther_ SVM	Precision	100	100	100	76	100	100	100
	Recall	86	91	87	100	93	100	92
	F1-score	92	95	99	88	96	100	96
Vi_ SVM	Precision	62	59	77	95	85	82	63
	Recall	71	61	73	84	92	76	69
	F1-score	67	60	75	89	89	79	66
Fu_ SVM +	Precision	92	85	98	95	94	87	68
	Recall	83	70	78	90	97	92	87
	F1-score	88	77	87	93	95	90	76

TABLE 8. Compare the results in Exp02.

Model	Ther_SVM	Vi_SVM	Fu_SVM+
Average accuracy	77.2	94.9	86.8

Although the accuracy estimation ratio of the Fu_SVM+ model is low compared to the thermal image, it is higher than that of the visible image. It has been proven that temperature and emotion are closely related. Using temperature improves estimated results on normal images.

2) YOLO MODEL

For the dataset labeled for the YOLO model, to evaluate this dataset we use YOLOv5 and YOLOv7 network models to perform emotional training and classification from visible images and thermal images, and then perform fusion from these two models (Fig.18).

In this experiment (Exp.03), the dataset is extracted a small part from the image database above, the dataset of images is shown in Table 9. We divide the data into 2 random parts: the training dataset and the test dataset.

After training the model, the system will give the results including a bounding box and a confidence score.

TABLE 9. The number of images of each emotion in the dataset.

EMOTION	TOTAL	TRAINING SET	VALID SET	TEST SET
An	961	576	193	192
Dis	518	300	100	118
Fe	680	405	135	140
Ha	1210	729	244	237
Ne	177	96	32	49
Sa	1450	870	290	290
Su	480	288	96	96
Total	5476	3266	1088	1122

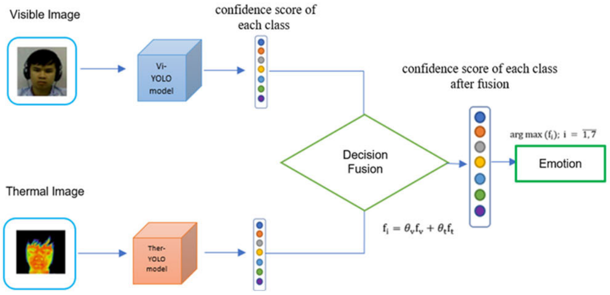

Visible Image
confidence score of each class
Vi-YOLO model
confidence score of each class after fusion
arg max (f$_i$); i = $\overline{1,7}$
Decision Fusion
Emotion
Thermal Image
Ther-YOLO model
f$_i$ = θ_vf$_v$ + θ_tf$_t$

FIGURE 18. Model YOLO for emotion estimation using visible and thermal images.

Since the training process of the YOLO model uses NMS (Non-max suppression) so it has filtered out the final predicted value, therefore, we have reduced the IOU index, from the default 0.45 to 0.25 to display the bounding boxes of the layers.

To determine the best emotion, we fuse two models trained from visible and thermal images using a linear formula with parameters specified based on the confidence score of each class.

With θv and θt as random weights, in this experiment, we use $\theta v = 0.4$, $\theta t = 0.6$, fv and f_t as confidence scores of each class. Finally, the emotion E is selected based on the maximum value of f_i.

$$E = arg\ max(f_i) \tag{1}$$

Table 10 shows the experimental results of the YOLO model.

D. ESTIMATION BY SPONTANEOUS INTENSITY EMOTION DATABASE (SIE)

In this section, a basic evaluation method is applied to validate the usability of the SIE database. Emotions Happiness, Sadness, and Anger had the most significant changes in facial temperature [6]. Therefore, in this study, we concentrate only on these three emotions for acquiring and estimating the different intensities.

TABLE 10. Experimental results on the YOLO model.

MODEL	PRECISION	RECALL	F1 SCORE	MAP@.5
Vi-YOLOv5	95.3	86.7	90.8	94.5
Ther-YOLOv5	87.6	79.2	83.1	87.6
Fusion-YOLOv5	89.9	81.4	85.4	89.7
Vi-YOLOv7	90.6	85.2	87.8	91.4
Ther-YOLOv7	97.8	87.7	92.5	95.9

The dataset has a capacity of 22.2 GB with 10 visible and 10 thermal videos of 10 participants respectively. The number of basic emotions included: happiness, anger, and sadness. In each emotion, there are 3 levels of expression in ascending order: "low, medium, high", and assign the corresponding value 1, 2, 3. Fig. 19, describes the emotion distribution of the number of frames in the SIE dataset with 3 levels for each emotion: happy, angry, and sad.

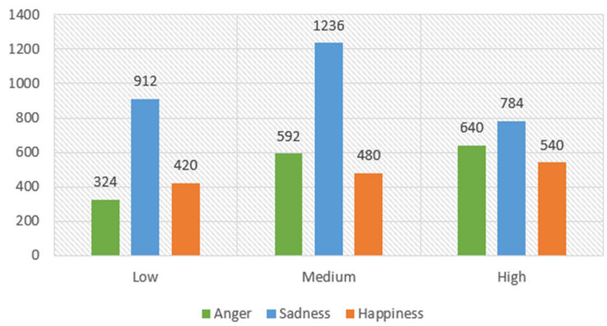


FIGURE 19. Frames distribution in the database (SIE).

The face in the thermal image is hard to detect, so we use the t-ROI to determine the region of interest. A ranking method proposed is a ranking framework (Fig. 20) with input data as the SIE dataset to estimate emotional intensity. The intensity values of SIE are discrete (low, medium, high), so the threshold model of the ranking learning approach is somewhat more optimal. Shashua and Levin [30] propose a ranking model that applies the principle of large margins combined with a threshold model. The solution of this approach is to find a set of parallel hyperplanes that divide the data in a certain monotonous order. Li and Lin [31] extend the above solution in the context of taking into account the sensitive cost value to improve the performance and convert the ordinal ranking problem into binary classification. Inheriting the results of [31], we apply the RED-SVM algorithm as a level estimation framework.

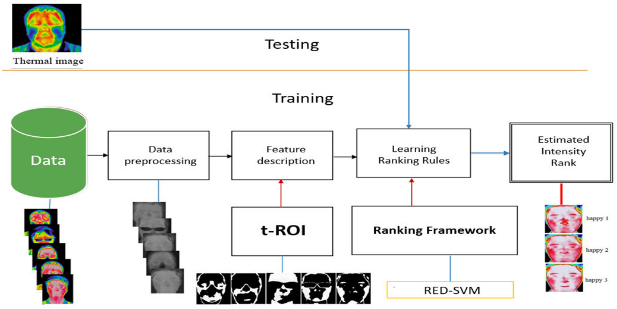


FIGURE 20. Overview of the model to estimate emotional intensity t-RankingSIE.

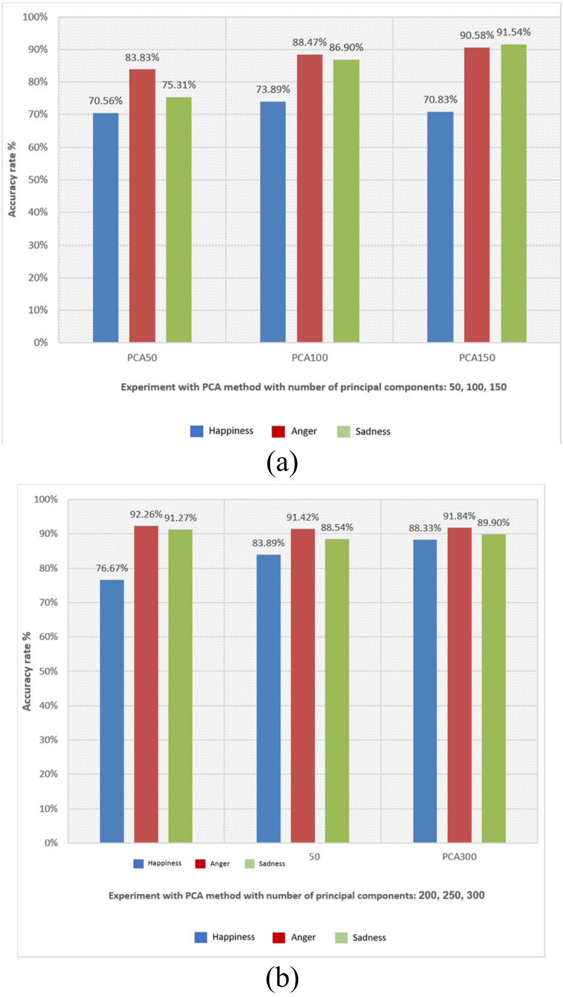
TABLE 11. Compare between datasets.

DATASET	KTFE	NIVE	KTFEv2
Centralization of objects	The objects in this dataset are many kinds. Thus, it does not tend to be a determined kind.	215 subjects (157 males and 58 females), ages 17 to 31	30 subjects (17 males and 13 females) including the age group from 12 to 32 years old from Vietnam, Japan, China, and Thailand
Multi-levels of emotions	Each emotion only has one level. Thus, they are not suitable with data in the real-world	Using visible and thermal images, there is only one facial expression database	Video recordings of human emotion in the visible and infrared spectrum, collected spontaneously.
Thermal information	Must use specialized copyrighted software that comes with the camera.	N/A	The database is stored by temperature files as CSV files, and it can be analyzed easily
Usability	The thermal file is a specific file that cannot be read from normal software It must use specialized copyrighted software that comes with the camera. Users are difficult to use it.	The dataset contains videos large-scale divided into small packages for download.	Easy to use, it contains the temperature matrix CSV file, the visible, and thermal images have been annotated

Experimental results of the t-RankingSIE model to estimate the emotional level made with dimensionality reduction by the PCA method have the number of main components in order: 50, 100, 150, 200, 250, and 300, shown in Fig. 21.

VI. DISCUSSION

There are many datasets for facial emotions. However, the collected data lacked natural spontaneous visible, and thermal videos. They do not notice the time lag phenomenon. Because emotion and expression are not synchronized, the time to stimulate emotion takes some time after expression occurs. This section compares the collected data and our dataset based on these criteria:



(a)

(b)

FIGURE 21. Experiment with PCA method with a number of principal components: 50, 100, 150; (b) Experiment with PCA method with a number of principal components: 200, 250, 300.

- **Centralization of objects:** In fact, each kind of object, such as children, teenager, young, middle age, older, will perform their emotions differently. if the objects of the dataset tend to a determined kind, the dataset will describe emotions for this object kind more clearly and it can be used to analyze discerningly.
- **Multi-levels of emotions:** An emotion has many levels. The dataset needs to be classified through those levels of emotions. From that, it can be effective to apply in the practice.
- **Thermal information:** A thermal camera can detect items since they are rarely exactly the same temperature as the objects around them, and they will appear as distinct in a thermal image. Hence, the dataset has to store information caused by heat or temperature which is useful to detect objects in the practice.

- **Usability:** This criterion is the completeness criterion for dataset requirements specification standards. The built dataset has to be used easily. It supports users to do their tasks fastly.

Table 11 compares current datasets and the built dataset in this study based on those criteria:

Despite our databases overcoming some limitations compared to other databases, KTFEv2 also has limitations that should be considered:

- KTFEv2 has a limited diversity of participants, which may reduce its generalizability and limit its ability to capture the full range of human emotions. The more participant, the better database.
- The diversity of participant control is not high enough. There is a lack of Europe and American participants.
- The education level is pupil and students only.

VII. CONCLUSION AND FUTURE WORK

In this research, the KTFEv2 database for expression and emotion recognition is proposed. This dataset includes both natural spontaneous visible and thermal IR videos. It is built based on standards of data collection and contains seven spontaneous emotions of thirty subjects. Besides, this dataset is spontaneous of human emotion with various intensities containing both visible and infrared videos; in which, each emotion is classified into three levels (low, medium, and high). It is also easy to use with the temperature matrix as a CSV file, the visible, and thermal images have been annotated.

KTFEv2 database is also compared with other databases based on applicability criteria. The built dataset has been analyzed by modern models of machine learning for emotion recognition, such as CNN, ResNet50, YOLO, and combined models. The acquired results are feasible and show that this dataset is effective to apply in the practice. The built database is more useful and potential to apply in the real-world. The database is freely available to the academic community and is easy to access available at there.[1]

In the future, the dataset will be continuously updated to develop, especially the thermal data for contributing better results. Besides, the dataset can be combined with facial features [32] as prediction rules of human emotions to be more useful in the real-world. Moreover, the technique of combining thermal infrared information can be applied in the processing of medical images [34], [35].

REFERENCES

[1] C. Darwin, Ed., *The Expression of the Emotions in Man and Animals*. London, U.K.: John Murray, 1872.

[2] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[3] A. Saxena, E. Y. K. Ng, and S. T. Lim, "Infrared (IR) thermography as a potential screening modality for carotid artery stenosis," *Comput. Biol. Med.*, vol. 113, Oct. 2019, Art. no. 103419.

[4] K. J. Havens and E. J. Sharp, *Thermal Imaging Techniques to Survey and Monitor Animals in the Wild: A Methodology*. New York, NY, USA: Academic, 2015.

[5] A. Kylili, P. A. Fokaides, P. Christou, and S. A. Kalogirou, "Infrared thermography (IRT) applications for building diagnostics: A review," *Appl. Energy*, vol. 134, pp. 531–549, Dec. 2014.

[6] M. M. Khan, R. D. Ward, and M. Ingleby, "Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature," *ACM Trans. Appl. Perception*, vol. 6, no. 1, pp. 1–22, Feb. 2009.

[7] R. Nakanishi and K. Imai-Matsumura, "Facial skin temperature decreases in infants with joyful expression," *Infant Behav. Develop.*, vol. 31, no. 1, pp. 137–144, Jan. 2008.

[8] S. Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 682–691, Nov. 2010.

[9] H. Nguyen, K. Kotani, F. Chen, and B. Le, "A thermal facial emotion database and its analysis," in *Image and Video Technology*, R. Klette, M. Rivera, and S. Satoh, Eds. Berlin, Germany: Springer, 2014, pp. 397–408.

[10] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.

[11] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 46–53.

[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[13] I. J. Goodfellow, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*. Berlin, Germany: Springer, 2013, pp. 117–124.

[14] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "DFEW: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2881–2889.

[15] K. Panetta, "A comprehensive database for benchmarking imaging systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 509–520, Mar. 2020.

[16] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[17] S. Li, L. Guo, and J. Liu, "Towards East Asian facial expression recognition in the real world: A new database and deep recognition baseline," *Sensors*, vol. 22, no. 21, p. 8089, Oct. 2022.

[18] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, and W. Zhang, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Inf. Fusion*, vols. 83–84, pp. 19–52, Jul. 2022.

[19] M. F. H. Siddiqui, P. Dhakal, X. Yang, and A. Y. Javaid, "A survey on databases for multimodal emotion recognition and an introduction to the VIRI (visible and InfraRed image) database," *Multimodal Technol. Interact.*, vol. 6, no. 6, p. 47, Jun. 2022.

[20] *NIST Equinox Database*. Accessed: Dec. 30, 2022. [Online]. Available: www.equinoxsensors.com/products/HID.html

[21] *IRIS Database*. Accessed: Dec. 30, 2022. [Online]. Available: https://vcipl-okstate.org/pbvs/bench/Data/02/download.html

[22] V. Espinosa-Duró, M. Faundez-Zanuy, and J. Mekyska, "A new face database simultaneously acquired in visible, near-infrared and thermal spectrums," *Cognit. Comput.*, vol. 5, no. 1, pp. 119–135, Mar. 2013.

[23] V. Espinosa-Duró, M. Faundez-Zanuy, J. Mekyska, and E. Monte-Moreno, "A criterion for analysis of different sensor combinations with an application to face biometrics," *Cognit. Comput.*, vol. 2, no. 3, pp. 135–141, Sep. 2010.

[24] R. Miezianko. (Jun. 2006). *Terravic Research Infrared Database*. [Online]. Available: http://www.cse.ohiostate.edu/otcbvs-bench/

[25] S. Wang, Z. Liu, Z. Wang, G. Wu, P. Shen, S. He, and X. Wang, "Analyses of a multimodal spontaneous facial expression database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 34–46, Jan. 2013.

[26] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cogn. Emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[1] https://drive.google.com/drive/folders/1GKx1hMJIkCWPAm923U_fMh9_G7MeTu63?usp=sharing

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] T. Nguyen, K. Tran, and H. Nguyen, "Towards thermal region of interest for human emotion estimation," in *Proc. 10th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2018, pp. 152–157.

[29] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Netw.*, vol. 22, nos. 5–6, pp. 544–557, 2009.

[30] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 1–8.

[31] L. Li and H. T. Lin, "Ordinal regression by extended binary classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 865–872.

[32] B. T. Nguyen, M. H. Trinh, T. V. Phan, and H. D. Nguyen, "An efficient real-time emotion detection using camera and facial landmarks," in *Proc. 7th Int. Conf. Inf. Sci. Technol. (ICIST)*, Apr. 2017, pp. 251–255.

[33] N. Nguyen, T. Nguyen, K. Tran, T. Luong, D. Nguyen, N. Vuong, P. Siritanawan, N. Tran, L. Nguyen, K. Kotani, H. Nguyen, L. Huynh, and H. Ho, "A spontaneous visible and thermal facial expression of human emotion database," in *Proc. 6th NAFOSTED Conf. Inf. Comput. Sci. (NICS)*, Dec. 2019, pp. 569–574.

[34] V. Pham, H. Nguyen, B. Pham, T. Nguyen, and H. Nguyen, "Robust engineering-based unified biomedical imaging framework for liver tumor segmentation," *Current Med. Imag.*, vol. 19, no. 1, pp. 37–45, Jan. 2023.

[35] N. Tran, H. Nguyen, N. Huynh, N. Tran, and L. Nguyen, "Segmentation on chest CT imaging in COVID-19 based on the improvement attention U-Net model," in *Proc. 21st Int. Conf. Intell. Softw. Methodolog., Tools, Techn. (SOMET)*, 022, pp. 596–606.

[36] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[37] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "FER-Net: Facial expression recognition using deep neural net," *Neural Comput. Appl.*, vol. 33, no. 15, pp. 9125–9136, Aug. 2021.

[38] F. Makhmudkhujaev, M. Abdullah-Al-Wadud, M. T. B. Iqbal, B. Ryu, and O. Chae, "Facial expression recognition with local prominent directional pattern," *Signal Process., Image Commun.*, vol. 74, pp. 1–12, May 2019.

[39] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan, "Learning a facial expression embedding disentangled from identity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6759–6768.

[40] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, and S. Wen, "SAANet: Siamese action-units attention network for improving dynamic facial expression recognition," *Neurocomputing*, vol. 413, pp. 145–157, Nov. 2020.

[41] Y. Fu, X. Wu, X. Li, Z. Pan, and D. Luo, "Semantic neighborhood-aware deep facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 6535–6548, 2020.

[42] M. Behzad, N. Vo, X. Li, and G. Zhao, "Towards reading beyond faces for sparsity-aware 3D/4D affect recognition," *Neurocomputing*, vol. 458, pp. 297–307, Oct. 2021.

[43] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7660–7669.

[44] N. Le, K. Nguyen, A. Nguyen, and B. Le, "Global-local attention for emotion recognition," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21625–21639, Dec. 2022.

HUNG NGUYEN received the B.S. degree (Hons.) in mathematics and informatics and the M.S. degree (Hons.) in computer science from the HCM University of Science, in 2003 and 2007, respectively, and the Ph.D. degree in computer vision from the Japan Advanced Institute of Science and Technology, Japan, in 2015.

From 2015 to 2017, he was a Research Associate at the Image Laboratory, JAIST. Since 2017, he has been the Dean of the Information Technology Faculty, Ho Chi Minh City University of Education. His research interests include computer vision, facial image analysis, robotics, affective computing, machine learning, and education.

Dr. Nguyen is a member of the SPIE and IAENG. He was recipient of the Best Paper Award from ICCSA 2014 and ICAEIC 2020.

NHA TRAN received the B.S. and M.S. degrees in computer science from the HCM University of Education, in 2014 and 2020, respectively.

He is currently a Lecturer with the Information Technology Faculty, Ho Chi Minh City University of Education. His research interests include computer vision, information retrieval, affective computing, and machine learning.

HIEN D. NGUYEN received the B.S. and M.S. degrees from the University of Sciences, VNU-HCM, Vietnam, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Information Technology, VNU-HCM, in 2020. He is currently a Senior Lecturer with the Faculty of Computer Science, University of Information Technology, VNU-HCM. His research interests include knowledge representation, automated reasoning, and knowledge engineering, especially intelligent systems in education, such as intelligent problem solvers. He received the Best Paper Award from SOMET 2022 and ICOCO 2022, the Best Student Paper Award from KSE 2020, the Best Presentation Clip Award from AI-Socha (Ho Chi Minh City), in 2020, and the Incentive Prizes of the Technological Creation Awards of Binh Duong Province, in 2021 and 2015.

LOAN NGUYEN received the B.S. degree from Foreign Trade University, in 2008, and the M.S. and Ph.D. degrees in knowledge management from the Japan Advanced Institute of Science and Technology (JAIST), Japan, in 2014 and 2017, respectively.

From 2008 to 2012, she worked at EY Ltd. Since 2017, she has been a Lecturer with the School of Accounting, University of Economics Ho Chi Minh City. Her research interests include knowledge management, internal audit, judgment and decision-making, and educational accounting.

Dr. Nguyen received the Best Poster Award from ECKM 2015 and the ISSIP-IBM Smart Service System Best Paper Award from HICSS50.

KAZUNORI KOTANI received the B.S. and M.S. degrees in electrical and electronic systems engineering and the Ph.D. degree in information science and control engineering from the Nagaoka University of Technology, in 1981, 1983, and 1990, respectively.

From 1983 to 1989, he was at the Consumer Products Research Center, Hitachi Ltd. He was a Research Associate at the Department of Electrical Engineering, Nagaoka University of Technology, from 1990 to 1991. He is currently a Professor with the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST). His research interests include computer vision, facial image analysis, and CG.

Prof. Kotani is a member of ITE and J-FACE.