Development and evaluation of an emotional lexicon system for young children

Whei-Jane Wei¹

Received: 8 February 2019 / Accepted: 30 March 2019 / Published online: 22 June 2019
© The Author(s) 2019

Abstract

Traditionally, children's emotion has been assessed by teachers according to observation. We should be able to detect children's emotions using algorithmic techniques. To achieve this goal, it is necessary to develop and evaluate an emotional lexicon based on the standardized test entitled Emotional Competencies Scale for Young Children (ECSYC). The purpose of this study was to establish the criterion-related validity. The methodology of this study was to firstly develop 40 scenarios based on ECSYC. Secondly, we developed the five-level criteria. Thirdly, this study implemented observer training and calculated inter-rater consistency reliability. Fourthly, observers categorized 200 children's replies into five levels. Fifthly, this study ranked the sequence of frequency of each level and completed the emotional lexicon. The findings showed that the Spearman's rho coefficient reached up to .406*. ($p = .026$), which is significant, indicating that Young Children Emotional Lexicon (YCEL) and ECSYC were significantly correlated. The accuracies of the emotion detection recognizer using a bimodal emotion recognition approach achieved 46.7%, 60.85% and 78.73% for facial expression recognition, speech recognition, and a bimodal emotion recognition, respectively. Findings confirmed that the YCEL is feasible for speech recognition. The bimodal emotion recognition accuracies increased 32.03% and 17.88% compared with using a single modal of facial expression recognition and speech recognition, respectively. It is feasible to automatically detect children's emotional development and bring the norm up to date.

1 Introduction

Emotions produce different physiological, behavioral and cognitive changes. Traditionally emotion has been indirectly assessed by teachers. However, teachers are too busy to assess children's emotions. Wei (2011) developed a standardized ECSYC comprised of 40 emotional competencies including four subscales of understanding self, understanding others, emotional adjustment, and self-motivation. Can children's emotional competencies be detected by child-computer interaction? Recently, the trends of big data and artificial intelligence have begun to save resources, including time, human power, and costs.

At present, the evaluation of children's emotional ability is assessed by teacher based on long-term observation. Most studies have also adopted indirect assessment by rating scales (Darling-Churchill and Lippman 2016). This

study directly detected children's emotional competencies based on expressions, voices, and drawing. The purpose of this study is to explore the correlation between emotional scores obtained by ECSYC and YCEL.

2 Literature review

2.1 ECSYC

ECSYC is a standardized test with the norm of emotional development of 4–5 years old children with high reliability and validity study. The reliability study has accomplished by stratified random sampling 1067 children in Taiwan. Cronbach's α has reached 0.98 and the validity study on Kaiser–Meyer–Olkin (KMO) has reached 0.951 (Wei 2011). It is a 5-level scale comprised four subscales and each subscale is composed of ten emotional competencies. It is assessed by teachers who understand children's emotion based on long-term observation.

✉ Whei-Jane Wei
wwj@go.utapei.edu.tw

¹ University of Taipei, 1, Aiguo West Road, Taipei 10048, Taiwan, Republic of China

2.2 YCEL

Visual-based emotion detection for “natural” human-robot interaction (HRI) creates an appropriate reaction according to the emotional detection of the communication partner (Udochukwu and He 2015). The rationale for developing the YCEL is to keep continuously update for 4–6 years old children’ emotional development average applying artificial intelligence with big-data analysis. To achieve this purpose, the researcher intends to develop a standardized YCEL. Researcher develops the TCEL based on ECSYC. Accordingly, children’s emotional competencies are assessed by 40 emotional theaters constructed by 40 emotional competencies of ECSYC. Finally, the reliability study on consistency is conducted by Cronbach’s alpha. Criterion-related validity is conducted by analyzing the correlation relationship between ECSYC and YCEL.

2.3 Emotional expression

A related study used emotional pictures for expression recognition and found that widely used facial emotion pictures (PoFA; i.e. “Ekman face”) and the Radboud Faces database (RaFD) are generally not considered to show real emotions (Dawel et al. 2017; Mehta et al. 2018). Another study used video to capture children’s facial expressions that the techniques can continuously detect the emotions each time unit determined by the researcher (Cunha 2018). In general, one second is most frequently used so that this study also uses one second as the time interval.

2.4 Emotional detection

Most existing emotion recognition techniques rely on the Convolutional Neural Networks (CNN) training model. Karen et al. proposed a VGG16 architecture based on the CNN network (Delplanque 2017). VGG16 emphasizes the importance of depth in CNN networks. Compared to CNN, each convolutional layer filter is compressed to 3×3 , replacing the larger filter commonly used in CNN. The efficiency of the convolution operation is improved, but the parameters in the fully connected layer of the last layer still account for 90% of the overall network. Therefore, Christian et al. proposed the Inception V3 architecture, which uses Global Average Pooling technology (Joseph and Strain 2003). Averaging the data of the pooling layer can solve the problem of having too many parameters and features captured by the layer.

Today’s CNN-based emotion recognition technology has solved the problem of having many parameters. However, if there are too many filters on the convolution operation layer, the time required for the convolution

operation is greatly increased. Thus, Francois et al. proposed the Xception architecture. Through depth-wise separable convolution technology, the speed of convolution operations is further improved and the amount of data is reduced (Daunic 2015). In the past, the general CNN convolution operation method convolved with each filter for each layer of data. Depth-wise separable convolution is a layer of data that is convoluted with only one filter, and a 1×1 filter is created for each output. Finally, it lets each layer of output data and each filter perform another convolution operation. In this way, the amount of computation of the convolution operation can be reduced from 1/8 to 1/9 of the general CNN.

A related study found that the success of emotional recognition depends on the transmission of information between the computer and the face (Mehta et al. 2018). The computer should be able to get information about the face instantly, such as emotions and gender. However, because the information of the face is highly complex, too many parameters arise in the process of machine learning. This makes it impossible to achieve instant recognition. Therefore, this study uses a real-time CNN-based emotion recognition program for identification. By using Global Average Pooling in the architecture, the information in each image is compressed while retaining important features. Moreover, the depth-wise separable convolutions technique reduces the amount of computation in the convolution operation, making emotional recognition of images possible in real-time.

Visual-based emotion detection for “natural” human-robot interaction (HRI) creates an appropriate reaction according to the emotional detection of the communication partner. HRI can be processed due to the development hardware and complex software applications (Strupp et al. 2008; Natarajan and Muthuswamy 2015). Dixit and Gaikwad (2016) used patch based face features and a SVM classifier to detect emotion through facial expressions. The experimentation carried out on The Japanese female facial expression database obtained 91% average accuracy for emotion detection of basic emotions, including happiness, anger, sadness, surprise, disgust, fear and neutral. The average emotion detection time was 1.1 s. It was a feasible follow-up research to identify which emotion could be easily detected. Vaish et al. (2019) maintained the highest accuracy possible while keeping the computational cost minimal. The system was tested on subjects that were not available in the dataset and gives a comparable result with other real-time emotion detection systems. In other words, some researchers have attempted to upgrade the accuracy of facial emotional detection. Accordingly, the study used Microsoft Azure system to detect facial expression instead of developing our own recognition system (Microsoft Azure 2018).

2.5 Statistical methods

The underlying methodology of previous studies is a lexicon-based method by which a lexicon is used to detect emotions in text based on speech recognition (Pajupuu et al. 2012; Kato et al. 2006). Using a single modal of speech recognition, the correct emotion recognition reaches 73% for happiness, 60% for angry, 55% for sadness, and the overall accuracy is 62% for all emotions (Hsu and Chen 2012). However, bimodal emotion recognition can reach an accuracy of 86.85%, an increase of 5% compared with using a single modal of emotion recognition (Song et al. 2015; Chuang and Wu 2004; Kessous et al. 2010). Furthermore, previous studies have indicated that it is impossible to achieve satisfactory results by recognizing emotions based on a single model for either speech or facial expression (Ma et al. 2019; Yang et al. 2017). Accordingly, this study applied a bimodal emotion recognition system by using both facial expression recognition (55%) and speech recognition (45%). Speech recognition included two parts. Part one which is scored according to lexicon, accounted for 38% since the time consumed for 80% of total time. Part two asks each child to draw and tell about the story. This part accounted for 7% because the time confused accounted for 20% of the total time. The reliability study was analyzed by examining the correlation between part one and part two.

This study developed an emotional lexicon by analyzing 200 children's answers to 40 questions from ECSYC. In addition to speech recognition, this study also applied automatic facial action analysis (Kapoor 2002) of expressions of seven emotions, including angry, despised, disgusted, happy, no opinion, sad and surprised. This study identified and calculated the frequency of seven emotions during each emotional theater game. According to the database, the score is converted into five levels multiplied by the number of occurrences, and finally added to the numerator. The sum of the occurrences of the seven emotions is the denominator, and the weight distribution of each emotion in each question is calculated. According to the five levels of the database, if the percentage reaches 80–100%, five points are given. This study used Microsoft's Project Oxford tool to establish expression recognition as the sub-criteria (Zhao et al. 2016).

In addition to facial expression detection, Udochukwu and He (2015) developed a rule-based approach to implicit emotional detection of text. The approach gave an average F-measure of 82.7% to "Happy", "Angry-Disgusted" and "Sad". Moreover, Cho et al. (2008) proposed a Bayesian-based method of emotion detection of voice. The method is based on the Bayesian networks which represent the dependence and strength between the dialogist's utterance

and his emotion. Darekar and Dhande (2017) used an algorithm in emotion detection to gain a higher accuracy level of speech analysis.

Karpouzis et al. (2007) applied an approach via facial, vocal, and bodily expressions recognition in emotional detection. They described a multi-cue, dynamic approach to detect emotion in naturalistic video sequences. Therefore, the hybrid approach proposed in this study also implemented facial, vocal, and language recognition.

3 Methods

This study attempts to answer the major research question: Is there a significant positive correlation between YCEL and ECSYC? To achieve this goal, the method of the study included in-depth interviews, focus group discussion, observation and experimental method. The implementation process included: (1) developing of emotional theater games based on 40 emotional scripts, (2) designing 40 questions for children to answer, and (3) developing five indicators for assessing each competency. For the experiments, (4) the researcher randomly selected 200 children aged 4–6 years, and (5) conducted 40 emotional theater experiments. After experiments, (6) the researcher executed observer training and carried out a reliability study of the consistency of four observers. (7) The validity study was undertaken by analyzing the correlation between ECSYC and YCEL.

This study uses the following three analysis techniques, Microsoft Azure Bing Speech-to-Text, Microsoft Azure Text Analytics technology, and Microsoft Azure Emotion technology. First, Microsoft Azure Bing Speech-to-Text is used to convert the children's answer to a sentence. The researcher used the Microsoft Azure Text Analytics technology to analyze the emotions contained in the text. The keywords of the sentence were compared to YCEL to determine the emotional score. This study then used Microsoft Azure Emotion technology to analyze the facial emotional changes throughout the process.

3.1 System design

This study established the YCEL for children and developed 40 emotional theater games. Young children identified facial expressions while enjoying the games and produced facial expressions (55% = A). Second, the teachers presenting the emotional theatre games asked the children questions and elicited responses. The scores were compared to YCEL according to the phonetic text of the children's answers, producing a speech recognition sentiment score (38% = B). Finally, we asked the children to draw the parts they watched, and verbally interpret and

convert the drawing in to text. The scores were then compared to YCEL to produce a character recognition sentiment score ($7\% = C$). The emotional scores obtained by young children ($A + B + C$) are related to a standardized test, ECSYC (standard), developed by the investigator. The purposes of the study are listed below: (1) to develop YCEL based on ECSYC (Wei 2011); (2) to detect emotion and score the competency using the following formula: facial expression recognition $\times 55\% +$ speech recognition scores $\times 38\% +$ explanatory drawing scores $\times 7\%$.

The study created a database to store the results of the Microsoft Azure Emotion analysis. The researcher created a QR Code maker as the ID for each participant. The ID records the participant's name, age, and gender, and the teacher's name. The database recorded the answer text (Speech-to-Text), text analysis results (TextAnalyticsResult) and emotional data obtained after analysis (FaceEmotion). The system architecture was constructed as shown Fig. 1.

In Fig. 1, the left hand column represents how a child interacts with the computer. The middle column shows the front-end interface design. The right column reveals the back-end database design.

3.2 Development of YCEL

1. Develop 40 emotional theater scripts based on ECSYC.
2. Determine the scoring standards of five rating scales for each emotional competency.
3. Conduct a focus group discussion with the kindergarten teachers to decide the scoring standards of five rating scales, as shown in Fig. 2.
4. Integrate the 40 emotional theater scripts into teaching activities and videotape the activities, as shown in Fig. 3.

Fig. 1 System architecture

Fig. 2 Focus group discussion

Fig. 3 Emotional theater is in processing

5. Undertake observer training using videos and establish inter-rater reliability (we reached a rate of 93.9%).
6. To develop YCEL, the researcher analyzes the 200 children's emotional competencies according to the videos and scoring standards.

3.3 Emotion detection model

The emotion detection software and the triangle verification method were applied (Zhang et al. 2018). The emotion detective technologies included Microsoft Azure Bing Speech-to-Text, Microsoft Azure Text Analytics, and Microsoft Azure Emotion API, as shown in Fig. 4.

The detailed steps of the emotion detection model are described as follows.

Fig. 4 Emotion detect model

Fig. 5 Emotional score calculation

3.3.1 Formula of emotional score

The formula of the emotional score equals $\text{FES} \times 55\% + \text{VTS} \times 38\% + \text{LES} \times 7\%$, as shown in Fig. 5.

Figure 5 indicates that the facial expression score accounted for 55% of a child's overall emotional score. The Microsoft Azure Bing Emotion API is very mature technology with an accuracy rate over 99% (Salvaris et al. 2018).

3.3.2 Facial expression score (FES) calculation

The expression emotion scores accounted for 55% of the children's overall emotional scores. Four children were analyzed by Microsoft Azure Emotion. After calculating the scores according to the formula, we calculated the number of emotions. Voice recognition turns the description into text and saves it to the database, as shown in Fig. 6.

According to the expression recognition software, analyze the emotion type per second, and finally calculate the total number of emotions. Divide the number of emotions by the total number of times and multiply by 100%. The formula is $(a/N) \times 100\% = A$, which is converted into a percentage. According to the fifth grade score (81–100% is 5 points), the expression recognition score is determined.

3.3.3 Voice to text score (VTS) calculation

Voice emotion score calculation accounted for 38% of the children's overall emotional scores. By using Microsoft

Fig. 6 Expression recognition score calculation method

Azure Text Analytics, each child's answer was analyzed using voice to text. It was impossible to manually compare the text with the vocabulary database, thus the application of AI in this study was essential for determining the VTS emotion score.

3.3.4 Language emotional score (LES) calculation

The language emotional score accounted for 7% of the children's overall emotional scores. Upon finishing each item, we asked the child to explain what he/she drew and then converted the voice to the text. We then compared it with the vocabulary database to get the language emotional score. In order to confirm that the data obtained meets the requirements, a pilot experiment was conducted.

3.4 Implementation of the pilot study

To verify the emotional scoring formula, the researcher implemented a pilot study to discover any problems.

3.4.1 Register ID

Each child's ID number was used to store the child's photo, drawing and database. Press Enter to jump to the photo taking step. Take a photo. The image is automatically captured and the photo is accessed based on the child's ID number. The child's photo and his teacher's picture will automatically jump to the screen.

In Fig. 7, the avatar shown on the screen represents the kindergarten teacher. The child image was shot by the camera and projected onto the screen.

3.4.2 Select the emotional theater game and start the video display

The program side plays the corresponding video by clicking the corresponding number. After the video is played, it automatically jumps to the system to ask questions.

Fig. 7 Teacher-child interaction

3.4.3 Set up the system and try the connection

The researcher installs the system background, microphone preparation and Google speech recognition system, and ensures the Wi-Fi is connected.

3.4.4 Implement child-computer interaction

When the system asks a question, the animation of the child–computer interaction is displayed on the screen. The voice recognition converts the text and compares it with the database. If the text is not in the database, it is saved and judged by machine learning. Finally, the teacher asks the child to draw what he watched and the system saves it as a jpg file. The teacher asks the child to describe he/she has drawn, as shown in Fig. 8. The children reply to the teacher's question in Fig. 9.

3.4.5 Light up the LED display according to facial expression score

At the end of each theater game, the system show a score representing the child's emotion. The LED display lights up accordingly.

4 Results

Table 1 shows that the α of YCEL is 0.75, and Table 2 shows that the α of ECSYC is 0.98. Table 4 shows that there is a significant correlation between YCEL and ECSYC, $r = 0.41$ ($p = 0.026^*$). The conclusions of the study suggest there is a significant positive correlation between YCEL and ECSYC. This study found that YCEL is suitable for machine measurement. A child's emotional development norm can be constantly updated automatically when YCEL is applied with machine learning.

Fig. 8 Teacher asks child to talk about his drawing

Fig. 9 The children reply to the teacher's question

4.1 Reliability analysis of YCEL

The reliability analysis evaluates the reliability of the whole scale. The reliability evaluation of this study is based on “developing a common emotional vocabulary database for children” to perform Cronbach's Alpha detection. The reliability of YCEL was estimated as follows:

Table 1 shows that the YCEL reliability coefficient Alpha was Cronbach's α 0.747, and the normalized reliability coefficient was 0.710. The normalized α indicates the influence of the unequal volatility of each subject, and the corrected coefficient.

Table 2 shows that the ECSYC reliability coefficient Alpha was Cronbach's α 0.976, and the normalized reliability coefficient was 0.976. The normalized α represents the influence of the unequal volatility of each subject, and the corrected coefficient.

4.2 Correlation analysis between YCEL and ECSYC

Table 3 shows that the sample averages of YCEL and ECSYC are 35.6 and 42.7, respectively.

Table 4 showed that the Spearman's rho coefficient reached up to 0.406*. ($p = 0.026$), which was significant, indicating that YCEL and ECSYC were significant. When it came to a significant correlation, ECSYC is a standardized test with high reliability and validity, so the YCEL developed in this study also has high reliability and validity. The findings suggest that applying YCEL is feasible for detecting emotional competency according to child-computer interaction in terms of emotional competency according to child-computer interaction in terms of emotional theater games.

Table 1 Reliability statistics of YCEL

Cronbach's Alpha value	Cronbach's Alpha value according to standard items	Number of items
0.747	0.710	37

Table 2 Reliability statistics of ECSYC

Cronbach's Alpha value	Cronbach's Alpha value according to standard items	Number of items
0.976	0.976	40

Table 3 Descriptive statistics

	Average value	Standard deviation	Number
ECSYC	35.6000	8.37731	30
YCEL	42.6667	6.13263	30

4.3 Database of children's frequently used emotional vocabulary

Figure 10 is an example of a lexicon of competency for YCEL (#1). The % represents the frequency of the lexicon at each level of the 5-level rating scale. There are five rating scales according to scoring standards, which were developed according to the following steps. First of all, this study developed 40 scenarios based on ECSYC. Secondly, we developed the five-level criteria that were categorized by kindergarten teachers. Thirdly, we implemented observer training and calculated inter-rater consistency reliability. Fourthly, the teacher asked children questions to reflect the children's emotional competencies. Fifthly, four observers categorized 200 children's replies into five levels. Sixthly, this study ranked the sequence of frequency for each level and completed the emotional lexicon.

4.4 Emotional detection by child–computer interaction

Results of a pilot study of emotion detection are summarized in Table 5 as follows.

5 Passed: Say two kinds of emotional words	4 Partially passed: Say one emotional words	3 Non-emotional word	2 Wrong answer	1 Didn't answer
Sad(25%) not very good/ Uncomfortable/ grieved / bad / worried / Unhappy(13%)	Happy(60%) Well(10%) joyful / a little happy/ Fulfilled/ Glad/ Good(6%)	Very uncomfortable(13%) Thank you(13%) stomachache(13%) Too much(13%) Want to go home to sleep(13%) Want to go to the toilet(13%) not good(13%) Crying(13%)	I feel very happy eating a lollipop	Did not answer

Fig. 10 An example of a lexicon for emotion competency (#1)

Table 5 is an example of the verification of the emotional detection formula: facial expression recognition scores $\times 55\%$ + speech recognition scores $\times 38\%$ + explaining drawing scores $\times 7\%$. The emotional competency scores of ECSYC are 3.59, 2.98, 1.44, and 3.56 for child #1, child #2, child #3, and child #4, respectively. ECSYC are assessed by teachers based on long-term observation and understanding each child.

As for the accuracy of facial expression recognition scores, child #1 achieved $(2/3.59) \times 100\% = 55.71\%$. Child #2, child #3, and child #4 achieved scores of 33.56%, 69.44%, and 28.09%, respectively. In sum, the average of the four children's facial expression recognition accuracies is 46.7%.

For the accuracies of speech recognition scores and drawing explanation scores, child #1 achieved $[(4 \times 38\% + 3 \times 7\%)/3.59] \times 100\% = 78.83\%$. The accuracies of speech recognition scores and drawing explanation scores were 60.40%, 40.97%, 63.2% for child #2, child #3, and child #4, respectively. In sum, the average

Table 4 Correlation analysis

			ECSYC	YCEL
Spearman's rho coefficient	ECSYC	r	1000	− 0.406 ^a
		Significance (two-tailed)		0.026 ^a
		N	30	30
	YCEL	r	− 0.406 ^a	1.000
		Significance (two-tailed)	0.026	
		N	30	30

^aThe relevant significance level is 0.05 (two-tailed)

Table 5 Emotional detection scores

Facial expression recognition detection dominated 55%										speech recognition dominated 45%	emotion score= facial expression recognition scores * 55%+ speech recognition scores * 38%+ explaining drawing scores * 7%.	
										Voice to Text Score: speech recognition scores dominated 38%	draw and tell about the story: explaining drawing scores dominated 7%
child	Sec.	Surprise	Sad	Angry	Afraid	Happy	Disgust	Doubt	Average	
1												
1												
Times												
%												
Score												
2												
1												
Times												
%												
Score												
3												
1												
Times												
%												
Score												
4												
1												
Times												
%												
Score												

speech recognition accuracy of the four children was 60.85%.

As a conclusion, the emotion detection accuracies of the emotion recognizer using a bimodal emotion recognition approach achieved 46.7%, 60.85% and 78.73% for facial expression recognition, speech recognition, and bimodal emotion recognition, respectively. Findings confirmed that the YCEL is feasible for speech recognition. The bimodal emotion recognition accuracies increased $32.03\% = (78.73\% - 46.7\%)$ and $17.88\% = (78.73\% - 60.85\%)$ compared with using a single modal of facial expression recognition and speech recognition, respectively.

5 Discussion

5.1 Emotional lexicon for children

The findings of this study could effectively improve the emotional ability of young children, and support the emotional theater experiments. The proposed approach offers a feasible and effective design to improve children's emotional ability. The results were same as those of related studies (Joseph and Strain 2003; Poventud et al. 2015). The teacher in the study told the children that the puppet was frustrated. The teacher asked the children to pay attention

to the teacher's story-telling and what the teacher said to the puppet. A related study also indicated that using seven effective strategies to improve children's vocabulary had a significant relationship with students' social emotional vocabulary scores (Daunic 2015).

Emotional recognition combining by facial expression and speech (Truong et al. 2007) has been applied in health monitoring and e-learning. There is a growing need for the development of agreed standards in automatic emotion recognition research. Therefore, this study developed five levels of assessing standards for each emotional competency. Also, based on big data, this study developed a database of emotional development with an incremental norm.

5.2 Correlation between YCEL and ECSYC

This study aims to develop YCEL based on the standardized test of ECSYC for autonomous emotion detection using a bimodal emotion recognition approach. The methodology of the bimodal approach is to combine speech recognition and facial expression recognition. In this design, scoring standards of five levels are programmed based on the emotional competencies of ECSYC. Using emotional theater games, the bimodal emotion recognition detection analysis showed that YCEL and ECSYC are

significantly correlated. Findings showed that Spearman's rho coefficient reached up to 0.406*. ($p = 0.026$), which is significant. ECSYC is a standardized test with a reliability and validity study published by Psychological Publishers in Taiwan. ECSYC is a criterion to support the YCEL, which is also a standardized test with high reliability and validity.

5.3 Emotion detection with bimodal emotion recognition

The process and rationale of the first purpose is to develop the criteria for each level of the 5-level rating scale. In this process, we developed 40 emotional games based on 40 emotion competencies. Each question reflected on each emotion competency. An example of the lexicon database is shown in Fig. 10.

The findings of the study are the same as those of a related study of bimodal emotion recognition accuracies; i.e., we reached increases of 12.15% and 3.54% compared with using a single modal of speech recognition and facial expression recognition, respectively (Song et al. 2015). The proposed method and architecture outperform other previously proposed approaches.

The system design of this study applied an open source method, in terms of real-time face emotion detection. Real-time image emotion recognition was executed by combining the OpenCV and CNN training models. The training dataset system used here is fer2013. CNN undergoes multiple convolutions, pooling, and full connection to determine the consequences of emotional expression.

The system compares the matching features of all the regions in the image at the beginning when the system does not know the characteristics of the feature. The mechanism used in this step is a convolution operation, where the degree of coincidence results is used to convolve the pixels of each block of a fixed size and the feature detector to obtain the block and the feature.

After the system repeats convolution operation and pooling, it enters the step of full connection. The system also transforms the pooled result matrix into one dimension and allows the values in the matrix to vote on the result. The result with the highest number of votes is the result of the current identification. The different values reflect varied degrees of discrimination. For example, some features better reflect the emotion of happiness. The number is expressed in terms of weight or connection strength.

Related studies have developed robots with emotional abilities similar to humans, giving them the ability to generate and express emotions (Mehrabian 2017). Based on a literature review, this study assumes that facial expression account for 55% of emotional scores. The voice to text recognition scores make up 38% of emotional scores, while the remaining 7% of emotional scores come

from language recognition scores. Therefore, it is very important to be able to correctly express the emotional expressions of the face. As long as a facial expression can be correctly expressed, most of the expression of a child's emotion can be identified.

Other studies have also pointed out that human expression involves the muscle movement of the face. A specific expression can be assigned to each action unit according to the intensity of the contraction. Therefore, by detecting the facial expression, we can observe the emotion of the subject. However, this method is very time consuming. Thus, a child-computer must be used to speed up the facial expression recognition process (Delplanque 2017).

6 Conclusion

ECSYC is a standardized emotion test with high reliability, high validity, and the average of emotional development for 4–6 years old children in Taiwan. It is assessed by children's teachers. Is it possible to develop a standardized automatic emotion recognition system from facial and speech recognition? It is a challenging task that relies on an emotional lexicon developed in this study.

To achieve this goal, this study firstly developed 40 child-computer emotional theater games based on the standardized emotional test of ECSYC. The study has demonstrated that it is feasible to automatically detect children's emotional development by child-computer emotional theater games. The contribution of this study is the accomplishment of the YCEL so that children's emotion can be automatically detected and it can bring the norm up to date. The follow-up study is in processing using deep learning to automatically discover emotionally relevant features. The average of emotional competency scores is automatically calculated and the norm is always kept in an updated version.

Acknowledgements This study was approved by the Ministry of Science and Technology and the research plan number is MOST 107-2221-E-845-004. I would also like to thank all those colleagues who participated and offered support. Without their assistance, this research project would not have been carried out smoothly. I would like to express thanks to the 3&3 International Education CORP and Taipei Private Sanmin Kindergarten for their assistance during the experiments.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Cho J, Kato S, Itoh H (2008) A biphasic-bayesian-based method of emotion detection from talking voice. In: The international conference on knowledge-based and intelligent information and engineering systems. Springer, Berlin
- Chuang ZJ, Wu CH (2004) Multi-modal emotion recognition from speech and text. *Comput Linguistics Chin Lang Process* 9(2): 45–62. <https://www.aclclp.org.tw/clclp/v9n2/v9n2a4.pdf>
- Cunha P (2018) README.md. GitHub, Inc. <https://github.com/haslab/Electrum/blob/master/README.md>
- Darekar RV, Dhande AP (2017) Toward improved performance of emotion detection: multimodal approach. In: ICDECT 2017. Springer, Singapore. https://doi.org/10.1007/978-981-10-1678-3_42
- Darling-Churchill KE, Lippman L (2016) Early childhood social and emotional development: advancing the field of measurement. *J Appl Devel Psy* 45:1–7
- Daunic AP (2015) Developing social-emotional vocabulary to support self-regulation for young children at risk for emotional and behavioral problems. *Int J Sch Cog Psychol*. <https://doi.org/10.4172/2469-9837.1000143>
- Dawel A, Wright L, Irons J, Dumbleton R, Palermo R, O’Kearney R, McKone E (2017) Perceived emotion genuineness: normative ratings for popular facial expression stimuli and the development of perceived-as-genuine and perceived-as-fake sets. *Behav Res Methods* 49(4):1539–1562. <https://doi.org/10.3758/s13428-016-0813-2>
- Delplanque S (2017) A comment on Prescott’s call for prudence and rigor when measuring emotions. *Food Qual Prefer* 62:372–373
- Dixit B, Gaikwad A (2016) Non verbal approach for emotion detection. In: Recent developments in intelligent information and database systems. Springer, Cham, pp 377–386. https://doi.org/10.1007/978-3-319-31277-4_33
- Hsu CY, Chen CP (2012) Speaker-dependent model interpolation for statistical emotional speech synthesis. *EURASIP J Audio Spe* 1:21
- Joseph GE, Strain PS (2003) Enhancing emotional vocabulary in young children. *You Except Child* 6(4):18–26
- Kapoor A (2002) Automatic facial action analysis. Diss. Massachusetts Institute of Technology. <https://vismod.media.mit.edu/tech-reports/TR-552.pdf>
- Karpouzis K, Caridakis G, Kessous L, Amir N, Raouzaoui A, Malatesta L, Kollias S (2007) Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. Artificial intelligence for human computing. Springer, Berlin, pp 91–112
- Kato S, Sugino Y, Itoh H (2006) A bayesian approach to emotion detection in dialogist’s voice for human robot interaction. In: Setchi Rossitza, Jordanov Ivan (eds) International conference on knowledge-based and intelligent information and engineering systems. Springer, Berlin, pp 961–968
- Kessous L, Castellano G, Caridakis G (2010) Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *J Multimodal User In* 3(1–2):33–48. <https://doi.org/10.1007/s12193-009-0025-5>
- Ma Y, Hao Y, Chen M, Chen J, Lu P, Košir A (2019) Audio–visual emotion fusion: a deep efficient weighted approach. *Sci Dir* 46:184–192. <https://doi.org/10.1016/j.inffus.2018.06.003>
- Mehrabian A (2017) Nonverbal communication. Routledge, New York
- Mehta D, Siddiqui MFH, Javaid AY (2018) Facial emotion recognition: a survey and real-world user experiences in mixed reality. *Sensors* 18(2):416
- Microsoft Azure (2018) Cognitive services try experience. <https://azure.microsoft.com/en-us/try/cognitive-services/>
- Natarajan P, Muthuswamy V (2015) Multi-view face expression recognition: a hybrid method. In: Suresh et al (eds) Artificial intelligence and evolutionary algorithms in engineering systems. Springer, New York, pp 799–808. https://doi.org/10.1007/978-81-322-2135-7_84
- Pajupuu H, Kerge K, Altrov R (2012) Lexicon-based detection of emotion in different types of texts: preliminary remarks. *Appl Linguistics* 8:171–184. <https://doi.org/10.5128/ERYa8.11>
- Poventud LS, Corbett NL, Daunic AP, Aydin B, Lane H (2015) Developing social-emotional vocabulary to support self-regulation for young children at risk for emotional and behavioral problems. *Int J Sch Cog Psychol*. <https://doi.org/10.4172/2469-9837>
- Salvaris M, Dean D, Tok WH (2018) Deep Learning with Azure: Building and deploying artificial intelligence solutions on the microsoft AI platform. Apress, Berkeley. <https://doi.org/10.1007/978-1-4842-3679-6>
- Song KT, Han MJ, Hsu JH, Hong JW, Chang FY (2015) Bimodal emotion recognition method and system utilizing a support vector machine. *Intell Serv Robot* 3(3):151–162
- Strupp S, Schmitz N, Berns K (2008) Visual-based emotion detection for natural man–machine interaction. Springer, Berlin, pp 356–363. https://doi.org/10.1007/978-3-540-85845-4_44
- Truong KP, Leeuwen DAV, Neerincx MA (2007) Unobtrusive multimodal emotion detection in adaptive interfaces: speech and facial expressions. In: Schmorow DD, Reeves LM (eds) Augmented cognition, HCII 2007, LNAI 4565, Springer-Verlag, Berlin, Heidelberg, pp 354–363. https://wwwhome.ewi.utwente.nl/~truongkp/pubs/2007_truong_et_al_unobtrusive_multimodal_emotion_detection_in_adaptive_interfaces_speech_and_facial_expressions.pdf
- Udochukwu O, He Y (2015) A rule-based approach to implicit emotion detection in text. 23rd International Öconference on applications of natural language to information systems Natural Language Data Bases NLDB. Springer, Switzerland, pp 197–203. https://doi.org/10.1007/978-3-319-19581-0_17
- Vaish A, Gupta S, Rathee N (2019) Enhancing emotion detection using metric learning approach. In: Saini S, Rishi R, Sandeep S (eds) Innovations in computer science and engineering. Springer, Singapore, pp 317–323. <https://doi.org/10.1007/978-981-13-7082-3>
- Wei WJ (2011) Emotional competency scale for young children. Psychological Publishing Co. Ltd, Taiwan
- Yang TH, Wu CH, Huang KY, Su MH (2017) Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals. *J Ambient Intell Hum Comput* 8(6):895–906. <https://doi.org/10.1007/s12652-016-0395-y>
- Zhang S, Zhang S, Huang T, Gao W (2018) Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Trans Multimedia* 20(6):1576–1590. <https://doi.org/10.1109/TMM.2017.2766843>
- Zhao M, Adib F, Katabi D (2016) Emotion recognition using wireless signals. *Commun ACM* 61(9):91–100. <https://doi.org/10.1145/3236621>

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.