Emotion Analysis Using Emotion Recognition Module Evolved by Genetic Programming

Rahadian YUSUF*, Ivan TANEV*, Katsunori SHIMOHARA*

(Received March 30, 2016)

This research is aimed at developing an intelligent agent able to recognize people’s emotions and self-evolve. One objective of the research is to develop an emotion recognition module. Several methods have been explored in order to improve the results while maintaining feasibility of real-time implementations for later stages. The emotion recognition module is evolved by using Genetic Programming, and several optimizations have been explored as well. Also, during the development phase, we encountered questions regarding the best time of the day (morning, afternoon, or late afternoon) to capture emotion and which emotions are easier to recognize. We decided to perform another experiment on several recognition modules evolving using Genetic Programming and analyzed the results in order to move a step closer to understanding the best time to gather data for emotions and the detailed accuracy of emotion recognition by our evolved modules.

Key words: evolutionary algorithms, genetic programming, genetic algorithm, emotion recognition, affective computing

1. Introduction

Many researchers have aimed to develop computer agents to achieve a better user experience during computer use, with one of the leading scientists being Rosalind Picard who coined the term Affective Computing1, 2). These computer agents are tasked with recognizing a user's request and interacting with the user. However, the user may possibly be influenced by a certain emotion, so the disruption of user experience in operating the computer might actually come from the user. For example, an angry user might become more irritated when prompted by out-of-place cheerful words, despite the agents just being programmed to be nice. With this in mind, it is interesting to explore the possibilities of computer agents that can not only recognize a user's emotions but also understand the impact of their feedback on the user's emotions as well.

The field of emotion recognition is related to affective computing, for which emotion recognition systems have been developed. Most of these studies have implemented similar methods such as using still-image data, generalization of human expressions, or the non-pervasive sensors such as electrocardiography (ECG) or electroencephalography (EEG) 3). Meanwhile, there are several gestures that usually relate to human emotion, such as nodding or head movements, and the stillness or the rapid movements of a whole body might relate to anticipation or anxiety. Still images have failed to capture these gestures, since temporal information is not stored in a single image. One of the scientists who use temporal information is Alex Pentland, who stated that activity level of a person related to their emotion 4).

Using non-pervasive sensors also might make a user uncomfortable and may influence the user’s actual emotion. In addition to this issue, a non-pervasive sensor is often not practical for real-world applications due to the special setup of the environment that might be needed.

Most researchers also use generalization of human expressions. However, our preliminary data suggested that individuals might express emotions differently. This might happen due to cultural background, demographics, or simply because each individual is unique.

Our research uses a pervasive sensor that utilizes a video stream as inputs (in order to capture several gestures and temporal information) and focuses on a single user only. Our previous reports on our research have focused on improving and refining our developed emotion recognition module as well as exploring different approaches to see the possibility of a self-evolving agent.

During these development processes, we encountered interesting questions: when is the best time to capture emotion, and which emotion is the most difficult to understand?

To answer these questions, we started to analyze the accuracy of our system for each emotion (not a general accuracy), and performed a set of new experiments to...
compare several data gathered at different times.

In this paper, we will report results of our explorations and experiments in developing the emotion recognition module and in analyzing the emotions of users.

2. User-Specific Approach and Related Works

2.1. Affective Computing and Emotion Recognition

While most researches so far use generalization of human expressions, we employ a quite different approach: a user-specific approach. First, we discuss why we should take this approach and how much originality it has through comparison with related works.

Rosalind Picard from Massachusetts Institute of Technology (MIT) \(^1\), \(^2\) used the terminology of Affective Computing and opined that a computer can improve its usability by understanding emotion, either by recognizing a user’s emotion or stimulate positive emotions in human-computer interactions. She states that emotion is fundamental to human experience, influencing cognition, perception, and everyday tasks such as learning, communications, and even rational decision-making. Moreover, she also stated that technologists have largely ignored emotion and created an often-frustrating experience for people, in part because affect has been misunderstood and hard to measure.

There are numerous signals that might be related to human emotion, such as facial expressions or other physiological signals. According to Alex Pentland \(^4\), \(^5\) human body movements are related to basic emotions in terms of activity level. They represent the eagerness of a person to interact with other people. There are also several other honest signals that can make people readable, such as repetitions or confirmations. John F. Cohn \(^5\) stated the lack of temporal analysis on the many affect or emotion recognition methods, which also similar with Alex Pentland’s. Our research implements the temporal information into the recognition system as well. We implement this idea by adding temporal information to our system, to detect the activity level.

There are also several other emotion-related studies that do not use facial expressions but other signals. Signals from EEG \(^3\) to detect brain signals or ECG \(^3\) to detect heart signals have become the main input of several studies, each with varying results and applications. For this research, we prefer pervasive sensor in the form of Microsoft Kinect, instead of the non-pervasive sensor such as EEG or ECG.

Research into emotion is of course not limited only to affective computing. Psychologists have been conducting more research into this intriguing subject. One such psychologist is Paul Ekman \(^6\), \(^7\), who investigated emotions from the point of view of psychology. He went to societies secluded or isolated from the rest of the world to perform experiments, and from the results he concluded that several basic facial expressions existed regardless of culture. Ekman used a photograph of actors making facial expressions and asked the people in those societies to guess the emotion expressed. He claimed that since many people in the secluded societies could guess the emotions correctly, then several kinds of emotions are expressed similarly and can be recognized regardless of culture. He adopted the Facial Action Coding System (FACS) that described facial expressions on the basis of movements of face muscles. This research, unlike Paul Ekman’s research, uses user’s own data instead of data from actors. A generalized sample data for training might fail to detect unique user’s feature.

One of the face models that can be used because of its low polygon count is CANDIDE-3 \(^8\), which has been used by several researchers studying facial expressions. This face model constructed a 3D model of a face, and the vertices are created so that it can be used to represent FACS. This research uses the CANDIDE-3 face model to represent captured facial expression.

A survey paper on affect recognition models covered much recent research on affect and emotion recognition methods \(^9\). It stated that there are no available benchmark data to compare the various methods fairly and also surveyed many researches that focus on emotion recognition based on photographs of actors or detecting images of a video by focusing on images in each frame. Further, it also stated that, in many of the studies, the available training data often consisted of actors making exaggerated expressions. This is inline with Paul Ekman’s research that uses data of actors performing expressions, which could be exaggerated. Meanwhile, our research uses user-specific recognition and training instead of data from several subjects, since we aim to capture user’s unique features by obtaining data for longer period of time, and as natural as possible.

To sum up, our research uses a different approach: a user-specific recognition module, using not only facial expressions but also gestures and activity level, which contain the temporal information.

2.2. Emotion Classifications

To recognize emotions, the first step is to classify the emotions themselves. There are several types of emotion classifications, such as the circumplex model by Russel \(^10\), Plutchik \(^11\) also proposed a model that, similar to that of Russel, categorized emotions in according with similarities.

2.3 Evolutionary Approach

Further, we are exploring the use of an evolutionary algorithm and probing the possibility of a self-evolving agent in this research.

Paul Ekman uses human expert to analyze the emotion of each data, and it took hours even for a single picture. The automated system to recognize several important features is needed in order to speed up the process, especially our system also implements temporal information to be analyzed, thus the use of human expert is not possible. Another matter to consider is that our system aim to recognize unique user-specific features, which means a pre-determined formula is not usable. These two reasons alone require the system to take an automated
heuristic approach to the huge search space consisting of a lot of parameters (with several FACS and the temporal information); therefore we implement an evolutionary methodology.

Another benefit of using evolutionary methodology is we also aim to explore the possibility of performing automated adaptation of the system. Implementation of pattern recognition algorithm might give a good result, given a specific set of training data; however, our system aims for an evolving system based on evolving set of data, as we do not use a fixed database but a growing database of a specific user. One method to facilitate this requirement is by using evolutionary algorithm.

Another reason in using evolutionary algorithm is to preparing a framework for exploring the possibility of adaptive system; that is evolving the system to a new user. Based on these requirements, especially the unknown formula for recognition, the huge search space, and preparing framework for an adaptive system capable of evolution, we use evolutionary methodology in this research.

3. Proposed System

In our research, we employ the circumplex model that uses both Valence (pleasantness or general mood) and Arousal (which is similar to activity level) to represent emotions \(^1\). Fig. 1 illustrates the classification of emotions based on Valence and Arousal.

![Emotion Classification](image)

Fig. 1. Categorization of emotion used in this research.

The four categories that we use in our research can be described as follows:
- Happiness (positive valence, positive arousal)
- Relaxed (positive valence, negative arousal)
- Sadness (negative valence, negative arousal)
- Anger (negative valence, positive arousal)

Due to the lack of benchmark data, our research uses our own experimental data, while emphasizing the variety of such data (e.g., using different seating positions and lighting conditions to simulate real-world conditions, and not setting the environment for gaining specific kinds of data) in order to improve the validity of data and experiments, as well as to investigate the robustness of our system. However, our research is still limited by the minimum and maximum lighting conditions required by the sensor.

3.1. System Design and Developments

Our research \(^1\) uses a pervasive sensor from Microsoft Kinect. Microsoft Kinect acts similarly to a webcam and can be used as a source of perceptions for the system.

Fig. 2 illustrates the structure of the proposed system of the intelligent agent.

![Proposed System Design](image)

Fig. 2. Proposed system design.

There are several reasons for choosing Microsoft Kinect as a sensor. The first is because it can be categorized as pervasive sensor. A non-pervasive sensor might influence a user's actual emotion, making the user more anxious or uneasy. Thus, using a pervasive sensor such as Microsoft Kinect might reduce the influence of an alien sensor on the user.

Another reason is that Microsoft Kinect is an off-the-shelf commodity, widely available as a practical tool for home personal computers (PC). It is not dedicated to a computer agent but is a general device that can be used as a webcam and a gaming device.

This paper will only discuss the Emotion Recognition Module and its use for analyzing emotion. However, details of the proposed system design can be found in Tables 1 and 2.

<table>
<thead>
<tr>
<th>Component</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>A single user who interacts with the system</td>
</tr>
<tr>
<td>Kinect Sensors</td>
<td>The main sensor of the system, consisting of several cameras and microphones</td>
</tr>
<tr>
<td>Affective Agent</td>
<td>The main module to manage and synchronize flows of data, including the module for machine learning</td>
</tr>
<tr>
<td>Database</td>
<td>Storage of features</td>
</tr>
<tr>
<td>User Interface</td>
<td>To interact with the user</td>
</tr>
</tbody>
</table>

Table 1. Components of proposed system.
Table 2. Main features of the affective agent.

<table>
<thead>
<tr>
<th>Category</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptions</td>
<td>Features extracted</td>
</tr>
<tr>
<td></td>
<td>Collected features from database</td>
</tr>
<tr>
<td>Actions</td>
<td>Features to be stored in database</td>
</tr>
<tr>
<td></td>
<td>Features to be deleted in database</td>
</tr>
<tr>
<td></td>
<td>Proper response</td>
</tr>
<tr>
<td>Behaviors</td>
<td>Synchronization</td>
</tr>
<tr>
<td></td>
<td>Recognition (classifier)</td>
</tr>
<tr>
<td></td>
<td>Self-evolution</td>
</tr>
<tr>
<td></td>
<td>Analyze</td>
</tr>
</tbody>
</table>

3.2. Feature Extraction

In our research, we selected several features from a user’s facial expressions based on the CANDIDE-3 face model used by Microsoft Kinect \(^{7,12-15}\). On the basis of this model, we selected nine Action Units (AU) that correspond to the detected facial movements: upper lip raiser, jaw lowerer, lip stretcher, brow lowerer, lip corner depressor, outer brow raiser, and three head tilt poses (yaw, pitch, and roll).

The raw data is obtained from Kinect in the form of a stream of video frames, which can be processed immediately to extract these nine features. The stream of raw data is fed into the system with a sampling period of 33ms.

Each AU is also normalized to a scale between -1,000 (no trace) and 1,000 (clear existence).

Fig. 3 below illustrates the 3D face model of CANDIDE-3 that is also used by Microsoft Kinect.

![Fig. 3. CANDIDE-3 Face Model \(^{7}\).](image)

Our ongoing experiments \(^{12-15}\) suggested that, as the extracted AUs only represent facial expressions, further feature extraction is needed. To extract activity level and temporal information, we need to know the changes of the AUs as well. Therefore, we put the extracted AUs into a buffer, and for every 100 rows (time-series data) of AUs, we extract three features for each AU:

- average value
- standard deviation
- power

Average value and standard deviation represent the general level and activity level of each AU, respectively. An integration or power of the signals is used as well as one of the extracted features. The total number of extracted features is 27 (9x3) for every 100 frames of AU extracted from Microsoft Kinect.

3.3. Genetic Programming and XGP

Our research aims to explore the feasibility of using genetic programming to evolve the Emotion Recognition Module. One of our main motivations is that using the genetic programming might enlighten us about the important features in recognizing emotions of a specific person, since we can analyze the evolved program.

XGP \(^{16,17}\) is an in-house GP engine that features XML-based genotype representations of candidate solutions (genetic programs), XML-schema that determines the allowed syntax of the genotypes, and a UDP channel to communicate between a fitness evaluator and the XGP. XGP manages the population of genetic programs and performs the main genetic operations – selection, crossover, and mutation.

As an in-house GP engine, XGP has been used intensively at our lab to evolve various classifiers and genetic programs. XGP also gave an advantage in the form of a faster development time due the versatility of usages: it only took a short time to change the XML-schema and adapt the desired syntax of genotypes for a program.

Evolving programs using XGP requires the engine as a Microsoft Windows application running in parallel with another application to evaluate the fitness of each individual sent by XGP. XGP will manage the population and GP operations.

Fig. 4 illustrates the common setup of training programs using XGP \(^{14,15}\).

![Fig 4. Common setup of training in XGP.](image)

3.4. Genotype Representation

In our work, we evolve (via XGP) the mathematical models (functions) that optimally recognize human emotions on the basis of the extracted facial features. These models (functions) are represented by XGP as parse trees featuring non-terminal (functional) and terminal symbols as elaborated in Table 3.
Our mathematical models (functions) represent the two axis of emotion, which are Valence (represented by the Valence tree) and Arousal (represented by the Arousal tree), as represented in Fig. 1.

Fig. 5 illustrates the tree structure of our mathematical model.

The positive or negative of each Arousal and Valence is determined on the basis of the result of the comparison of the sub-trees using comparison operators. A 'true' result is considered 'positive', while a 'false' result is considered 'negative'.

Each parse tree is represented using the following Backus–Naur Form (BNF), along with the syntaxes, terminal sets, and functions:

\[
\text{IF } "F" \text{ "Comp" } "F" \text{ THEN True} \\
\text{Comp:} = "<"|">" \\
F::= "Const"|"Var"|"F","Op","Const"|"Var"|"F" \\
\text{Const:} = "1"|"..10" \\
\text{Var:} = "v_0"|.."v_26" \\
\text{Op:} = "*"|"-="|"*="|"*"/
\]

### 3.5. Fitness Function

The fitness function determines the mechanism of evaluating the quality of each individual in XGP (i.e., the evolved mathematical function for emotion recognition). In our work, we use the Matthews Correlation Coefficient (MCC) \(^{18}\) as a fitness functions on the basis of its accuracy and precise predictions of valence and arousal of each data sample during training. We opted for this value instead of a simple accuracy calculation, because our system should consider both precision and recall as a single value, and MCC offers this possibility.

MCC can be expressed as the following equation.

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]

where:

- \(TP\) = True Positive
- \(TN\) = True Negative
- \(FP\) = False Positive
- \(FN\) = False Negative

MCC results in a value between -1 (exact opposite) to 1 (perfect match). The value of zero means total randomness. Our system scaled this MCC to fit within the range between 1 (perfect match) and 9,999 (exact opposite). The value of 5,000 indicates a total randomness.

### 4. Results and Discussions

Our previous experiments' objective \(^{12-15, 19}\) was to improve the accuracy of emotion classification on test data set while maintaining a reasonable evolution time. There was usually a trade-off between accuracy and time. The details of each report are as follows.

At first, our initial experiments \(^{12}\) suggested that different people might express emotions differently on the basis of their facial expressions and gestures. Several experiments showed that evolving the trees (represented in Fig. 5 separately into Valence tree and Arousal tree) improves the accuracy in test data sets \(^{13,14}\), with our results suggesting that several features might be related to a specific branch only (Valence or Arousal). By separating the evolution, we prevented the genes of one tree to crossover with another tree (e.g. from Arousal tree to crossover with gene from Valence tree, vice versa).

We improved the accuracy further by implementing a voting system by evolving multiple classifiers \(^{14}\). However, training time has significantly increased, which we think will reduce the possibility of an applicable self-evolving agent.

To have a self-evolving agent, we need to reduce the training time, thus we explored the possibility of evolving the voting system using a genetic algorithm \(^{15}\). The result suggested a trade-off, but a Pareto front between the time needed to train and accuracy in test data sets can be searched for.

To add another topic, we performed experiments to analyze several matters related to emotion and affective computing \(^{19}\); we wanted to analyze the effect of data acquisition during specific times. We categorized the times of data acquisition into morning, afternoon, and late afternoon.

Another objective of the experiment is to analyze whether there are differences in difficulties of learning different emotions. Therefore, unlike previous experiments, we analyze the accuracy of each emotion category (Happy, Sad, Relaxed, and Angry) instead of general performances.

As we focus on comparing different data sets for
evolving classifiers, we conducted the experiment by using a similar setup to our previous experiments.

4.1. Experiment Terminologies

We used several experiment terminologies on every experiments performed.

- **Genotype**: a single gene that formed the program (function).
- **Individual**: a single program resulting from XGP runs, in the form of XML, consisting of information regarding the tree structure that represents the program.
- **Generation**: One generation of genetic programming runs. In our experiments, we use 100 individuals per generation.
- **Sessions**: One set of XGP runs until termination criterion is reached. The termination criterion is one of the following: number of generations (100 is the maximum), fitness value (2, or near perfect match), or stagnation period (30 generations of stagnation).
- **Batch**: One batch of XGP is a number of sessions with the same purposes, goals, and settings, in order to achieve statistical data that can show the computational performance and robustness of the system.
- **Experiment**: One experiment consists of several batches (each with different setting) to be compared to investigate the strengths and weaknesses of each method. In the current experiment, the different setting is the data set used for training and testing.

4.2. Data Acquisition

The experiments performed three acquisitions of data from a single subject. The subject was required to imagine a situation that would make him show a specific emotion, and then the Kinect captured a video (and extracted the data) for around 1 minute per emotion.

The first, second, and third acquisitions were conducted in the morning (around 9 AM), afternoon (around 1 PM), and late afternoon (around 5 PM). The data acquired are labeled as Morning, Afternoon, and Late, respectively.

We ran four batches for each data set (12 batches in total). Each batch consisted of 15 sessions. From each batch, the five best individuals (in terms of Fitness Function) were selected. They were then used as classifiers with the implementation of voting (majority selection).

4.3. Experimental Results and Discussions

Tables 4, 5, and 6 show the accuracy in data sets of Morning, Afternoon, and Late, respectively. Shaded columns show where the data sets used for testing and training (or evolving classifiers) were the same.

From the tables, we can see that the average results for all emotions do not significantly differ; averages for the Morning, Afternoon, and Late data sets are 83.83%, 81.25%, and 80.67%, respectively. However, there are several major inaccuracies, especially when Late data was used for training and Morning data for testing, especially for recognizing when the subject was relaxed.

The results suggested that there is no significant difference in general performances between taking the data in the morning, in the afternoon, or in late afternoon; however, there are significant differences in that some emotions are not recognized well if we acquire data of a specific emotion at a specific time.

There might be several possibilities regarding this result. First, the data might have been incomplete. In some cases, the data might also have been insufficient, thus the evolved programs actually may not have learned enough.

Another possibility is that the stress levels (valence) and the arousal levels are dynamic and might have influenced the subject of the experiment to express his emotions differently; his expressions may have been influenced by his 'real' feelings (such as tiredness in the late afternoon, or sleepiness in the morning) at the moment of data acquisition. This can be examined by acquiring more data on different days, which can be done in our next experiments.

To illustrate the differences between the results, Fig. 6, Fig. 7, and Fig. 8 show graphs of accuracy using the training data of Morning, Afternoon, and Late, respectively.
4.3. Experimental Results and Discussions

We ran four batches for each data set (12 batches in total). The data acquired are labeled as Morning, Afternoon, and Late, respectively. The first, second, and third acquisitions were conducted in the morning (around 9 AM), afternoon (around 1 PM), and late afternoon (around 5 PM). The data extracted the data) for around 1 minute per emotion. To illustrate the differences between the results, Fig. 6, 7, and 8 show the test accuracy for each emotion across the morning, afternoon, and late batches, respectively.

Table 5. Test Accuracy in afternoon data set.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Morning Test</th>
<th>Afternoon Test</th>
<th>Late Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>90.00</td>
<td>95.50</td>
<td>90.00</td>
</tr>
<tr>
<td>Angry</td>
<td>80.00</td>
<td>85.50</td>
<td>80.00</td>
</tr>
<tr>
<td>Relaxed</td>
<td>70.00</td>
<td>75.50</td>
<td>70.00</td>
</tr>
<tr>
<td>Sad</td>
<td>60.00</td>
<td>65.50</td>
<td>60.00</td>
</tr>
</tbody>
</table>

Table 7. Average and standard deviation of each emotion.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Accuracy (%)</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>92.22</td>
<td>6.27</td>
</tr>
<tr>
<td>Relaxed</td>
<td>75.36</td>
<td>22.26</td>
</tr>
<tr>
<td>Angry</td>
<td>86.17</td>
<td>12.60</td>
</tr>
<tr>
<td>Sad</td>
<td>75.33</td>
<td>21.90</td>
</tr>
</tbody>
</table>

The experimental results shown in Fig. 9 suggest that happiness is the easiest emotion to recognize; it has not only a better average accuracy in tests but also robust and stable results (a low standard deviation). Relaxation and sadness are not only difficult to recognize but also have very volatile results (high standard deviations).

4.4. Discussions on Easily Recognized Emotions

By re-arranging the table, we can obtain a better representation of the results, so we can easily see which emotions are easier to recognize. Table 7 shows the averages and standard deviations of all tests from the experiment.

The reason behind this might be related to the tendency of people to try to hide their sadness (as sadness can be viewed as weakness), and that people might express anger or happiness with the intention of manipulating others (such as showing aggressiveness). These behaviors are quite similar to several phenomena in the natural world such as [Handicap principle].

Another possibility that might be related to the results is that sadness and depression are difficult to recognize, since people are becoming more adept at hiding and not showing signs of them to the world.

However, a more thorough analysis from the point of view of psychology might be needed.

5. Conclusion

Our research focuses on the evolution of an agent able to recognize emotions of a single user. Several approaches such as separating a tree structure for evolution, implementing a voting system, and evolving the voting system, showed improvements in accuracies and faster training time (which is required for a self-evolving agent).

During the development of an emotion recognition module, a question was raised regarding the best time of
the day for recognizing emotion. Using available data, we performed an experiment to evolve several classifiers using different training data and cross-tested the classifiers using each set of data. The controlled variable is that each data set is taken at different times.

The results from the experiment suggested that although the time of data acquisition might not make average accuracy for all emotions significantly different (around 80% to 83%), the accuracy for each emotion might vary.

Further, our experimental results also suggested that high arousal emotions (Happy and Angry) are easier to recognize. This might be related to the psychological reasons behind those emotions; showing anger or happiness might be a means for people to manipulate other people, such as to threaten a foe, prepare to fight, or inform allies of a current situation.

For future works, new experiments can be performed to clarify these results by collecting more data over several days. Also, an experiment on many subjects can also be done to see the robustness and generality. Adding more test cases is highly crucial for testing the robustness of the system.

References

6) P. Ekman, Emotions Revealed, (Times books, New York, USA, 2003).